DOCUMENT RESUME

ED 478 425 TM 035 109

AUTHOR	Stancavage, Frances B.
TITLE	An Agenda for NAEP Validity Research: NAEP Validity Studies. Working Paper Series.
INSTITUTION	National Center for Education Statistics (ED), Washington, DC.
SPONS AGENCY	National Assessment of Educational Progress, Princeton, NJ.
REPORT NO	NCES-WP-2003-007
PUB DATE	2003-04-00
NOTE	110p.
AVAILABLE FROM	Available only online: http://nces.ed.gov/pubsearch/ .
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	EDRS Price MF01/PC05 Plus Postage.
DESCRIPTORS	*Agenda Setting; *Data Analysis; Elementary Secondary Education; National Surveys; *Research Needs; Sampling; *Validity
IDENTIFIERS	*National Assessment of Educational Progress

ABSTRACT

The National Assessment of Educational Progress (NAEP) Validity Studies Panel undertook a systematic analysis to consider the domain of validity threats to NAEP and to identify the most urgent research priorities. A framework of six broad categories was developed: (1) the constructs measured within each of NAEP's subject domains; (2) the manner in which these constructs are measured; (3) the representation of the population; (4) the analyses of data; (5) the reporting and use of NAEP results; and (6) the assessment of trends. Panel subcommittees prepared papers laying out the critical validity issues in each area. These papers, which are presented in chapters 2 through 7 of this report, are: (1) "Validity Issues Representing Populations" (Donald H. McLaughlin, Peter Behuniak, and James R. Chromy) (Chapter 4); (2) "Issues and Recommendations on NAEP Data Analysis" (R. Darrell Bock, Albert Beaton, and Gerald DeMauro) (Chapter 5); (3) "Validity and Utility Issues in NAEP Reporting and Data Releases" (Frances B. Stancavage and Ina V. S. Mullins) (Chapter 6); and (4) "Estimating Trends from NAEP Scores: Rationale and Research Directions" (David Grissmer, Albert E. Beaton, and Larry Hedges) (Chapter 7). The panel reviewed these papers and set priorities in each area by a consensus process. Sixteen suggested studies or areas of study were identified. Four stood out as essential, nine others were considered "highly needed," and three were found to be less important. The panel indicated unanimously that studies are essential to evaluate the validity aspects of NAEP's new role under the No Child Left Behind Act. The appendix to chapter 4 is attached. (Contains 6 figures and 26 tables.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 478 425

NATIONAL CENTER FOR EDUCATION STATISTICS

Working Paper Series

NAEP VALIDITY STUDIES: AN AGENDA FOR NAEP VALIDITY RESEARCH

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Working Paper No. 2003-07 April 2003

Contact: Patricia Dabbs
Field Operations and Publication Management
Patricia.Dabbs@ed.gov

TM035109

U. S. Department of Education
Institute of Education Sciences

BEST COPY AVAILABLE

U.S. Department of Education
Rod Paige
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Statistics
Val Plisko
Associate Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high quality data to the U.S. Department of Education, the Congress, the states, other education policymakers, practitioners, data users, and the general public.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to:

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006–5651

March 2003

The NCES World Wide Web Home Page address is *http://nces.ed.gov*
The NCES World Wide Web Electronic Catalog is: *http://nces.ed.gov/pubsearch*

Suggested Citation

U.S. Department of Education, National Center for Education Statistics. *NAEP Validity Studies: An Agenda for NAEP Validity Research*, NCES 2003–07, by Frances B. Stancavage, et.al. Project Officer: Patricia Dabbs. Washington, DC: 2003

For ordering information on this report, write:

U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794–1398

Or call toll free 1–877–4ED–Pubs

Content Contact:
Patricia Dabbs
(202) 502–7332
Patricia.Dabbs@ed.gov

Foreword

In addition to official NCES publications, NCES staff and individuals commissioned by NCES produce preliminary research reports that include analyses of survey results, and presentations of technical, methodological, and statistical evaluation issues.

The ***Working Paper Series*** was initiated to promote the sharing of the valuable work experience and knowledge reflected in these preliminary reports. These reports are viewed as works in progress, and have not undergone a rigorous review for consistency with NCES Statistical Standards prior to inclusion in the Working Paper Series.

Copies of Working Papers can be downloaded as pdf files from the NCES Electronic Catalog (http://nces.ed.gov/pubsearch/), or contact Sheilah Jupiter at (202) 502–7363, e-mail: sheilah.jupiter@ed.gov, or mail: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 1990 K Street NW, Room 9048, Washington, DC 20006.

Marilyn M. Seastrom
Chief Mathematical Statistician
Statistical Standards Program

Ralph Lee
Mathematical Statistician
Statistical Standards Program

An Agenda for NAEP Validity Research

Frances B. Stancavage, *Project Director, American Institutes for Research*
Albert E. Beaton, *Boston College*
Peter Behuniak, *Connecticut*
R. Darrell Bock, *University of Chicago*
George W. Bohrnstedt, *American Institutes for Research*
Audrey Champagne, *SUNY, Albany*
James R. Chromy, *Research Triangle Institute*
Susan Cole, *American Institutes for Research*
Gerald DeMauro, *Office of State Assessment, New York*
Richard P. Duran, *University of California, Berkeley*
David Grissmer, *RAND*
Larry Hedges, *University of Chicago*
Gerunda Hughes, *Howard University*
Donald H. McLaughlin, *American Institutes for Research*
Ina V.S. Mullis, *Boston College*
P. David Pearson, *University of California, Berkeley*
Lorrie Shepard, *University of Colorado*

Developed by the NAEP Validity Studies (NVS) Panel
October 2002

The NAEP Validity Studies Panel was formed by the American Institutes for Research under contract with the National Center for Education Statistics. Points of view or opinions expressed in this paper do not necessarily represent the official positions of the U. S. Department of Education or the American Institutes for Research.

Table of Contents

Executive Summary i

Chapter 1. Introduction 1

Chapter 2. Subject Domain: What Is Being Measured? 5

Chapter 3. Subject Domain: How Is It Being Measured? 13

Chapter 4. Validity Issues: Representing Populations 21

Chapter 5. Issues and Recommendations on NAEP Data Analysis 31

Chapter 6. Validity and Utility Issues in NAEP Reporting and Data Releases 43

Chapter 7. Estimating Trends from NAEP Scores: Rationale and Research Directions 47

Chapter 8. Synthesis: An Agenda for NAEP Validity Research 61

Appendix to Chapter 4 .... A-1

Executive Summary

At the beginning of the 21st century, the National Assessment of Educational Progress (NAEP) continues to stand as a unique, and uniquely valuable, resource for American education. It is the only periodic measure of student achievement based on national probability samples, and it is the only method by which states can validly compare the academic progress of their students against common high standards. With the passage of Public Law 107-110 (No Child Left Behind), NAEP is expected to play a greater role in helping states judge the adequacy of their yearly progress, both overall and for important subgroups of students.

For over 30 years, while functioning as "the Nation's Report Card," NAEP has also maintained a level of methodological rigor that has served as a standard for other testing programs. This rigor is not without costs, however. Each year, NAEP faces new psychometric issues as it attempts to provide useful information to various audiences while responding to the changing educational and social contexts of testing. In this circumstance, continuous vigilance is necessary to ensure that NAEP results remain valid.

Since 1996, the NAEP Validity Studies (NVS) Panel has joined with the National Center for Education Statistic (NCES), the National Assessment Governing Board (NAGB), and the NAEP contractors to consider issues of NAEP validity. Over its tenure, the panel has published a number of studies on aspects of NAEP development and implementation, and it has helped ensure that the NAEP program will respond successfully to new demands and will not unintentionally compromise the integrity and rigor of its reports. Now the panel addresses NAEP's need for a comprehensive agenda for validity research.

Aspects of Validity

Validity is the extent to which the messages in NAEP reports accurately communicate the state of educational progress in America to educators, policymakers, and the public. If NAEP reports, for example, that the gap in student achievement in mathematics or reading is widening, many questions can be raised about the meaning of that report. Does the mathematics or reading tested by NAEP represent the kinds of mathematics or reading skills that are important for students to achieve? Is the way that NAEP tests these skills fair and accurate? Do the results represent the full student population? Do the analytical procedures accurately tie the data to general statements about achievement? Are the results stated sufficiently clearly and unambiguously in the report so that misinterpretations are avoided? Do the reported data adequately capture the trajectory of academic progress in these subject areas, that is, trends over time? These are the types of questions that speak to NAEP's validity, and validity research consists of studies undertaken to address these questions.

To prepare a systematic analysis of the domain of validity threats and to identify the most urgent validity research priorities, the NVS panel created a broad framework that encompasses all aspects of NAEP. The framework was broken down into six broad categories:

- The constructs measured within each of NAEP's subject domains
- The manner in which these constructs are measured
- The representation of the population
- The analysis of data
- The reporting and use of NAEP results
- The assessment of trends

A subcommittee of two or three panel members assumed responsibility for one of the six broad areas of validity concerns. Over the course of two panel meetings, each subcommittee (or one key author from the subcommittee) prepared a paper laying out the critical validity issues in its area and proposing studies to address these issues. The papers, which were subsequently revised after discussion by the full panel, are presented in chapters 2 through 7 of this report.

Finally, at a third panel meeting in November 2001, the panel focused on reaching a consensus on the priorities of different areas of validity research across the six broad categories. Sixteen studies, or areas of study, were derived from the subcommittee reports and rated by the full panel.

The importance of validity research can be evaluated in terms of the potential harm that will result if the research is not done and it is later discovered that a hypothetical threat to validity is not merely hypothetical. The studies were therefore rated on a 5-point scale (Essential, High, Moderate, Low, or Not Needed) that combines judgments about both the perceived seriousness of a problem should it occur and the likelihood of its occurrence. "Essential" conveys the NVS panel member's opinion that if a study of the type described is not undertaken, NAEP will surely be subject to potentially damaging criticism. "High" conveys the opinion that failing to undertake such a study would be a major gamble; "Low" conveys the opinion that NAEP could withstand criticism because the error is unlikely to occur or its impact is unlikely to be severe. "Moderate," of course, conveys an opinion between "High" and "Low." "Not Needed" conveys a concern that NAEP might be criticized for conducting such a validity study.

The ratings, it must be emphasized, indicate the level of need for studies in particular areas, not the details of specific studies. The NVS panel recognizes that some of the studies described are already being carried out or are being considered by others. Moreover, the panel makes no statement here about either particular study designs or authority to execute the studies. Rather, the priority ratings reflect more general perceptions of the need for NAEP to do these studies.

Study Priorities

The judgments of individual panel members were averaged, and the aggregate priorities for each validity study recommended by the NVS panel are shown in Table 1. The table also indicates the broad category of validity concerns from which each specific study was drawn. Studies related to "Uses," for example, are discussed in chapter 6, Validity and

Utility Issues in NAEP Reporting and Data Releases, while studies related to "Trends" are discussed in chapter 7.

Table 1–Aggregate NAEP Validity Research Priority Judgments

Validation Aspect	Study	Average Priority Rating
	Essential	
Uses	Meaning of "confirming state results"	4.00
Uses	Limits on NAEP's capacity to evaluate state results	4.00
Trends	Bridge studies	4.00
Construct	Alignment with state standards	3.79
	High to Essential	
Analysis	Estimation of item domain sampling error	3.57
Measurement	Accommodations	3.54
	High	
Measurement	Contaminations	3.25
Sampling	Representing excluded SD and LEP students	3.14
Construct	Definition: What is being measured	3.07
Sampling	Combined studies of population bias	3.00
Analysis	Direct estimation with minimal conditioning	3.00
Trends	Estimation of multi-time-point trends	3.00
Construct	Definition: What students do on the test	2.86
	Not High	
Uses	Evaluation of audience interpretations	2.54
Construct	Definition: Comparison with curriculum	2.43
Uses	Controls and supports for secondary analysis	1.08

Note: Averages are based on scaling panelists' responses from 0 (Not Needed) and 1 (Low) to 3 (High) and 4 (Essential).

Four studies stand out as essential and two received ratings that place them between essential and highly important. Among the remainder, seven studies are evaluated as highly important, and three are of lesser importance or problematic in some way.

Essential Studies

The NVS panel indicated unanimously that studies are "Essential" to evaluate the validity aspects of of NAEP's new role under P.L. 107-110, however that role is eventually operationalized. The panel was careful to distinguish between definition and validation. While recognizing that the need for defining the concept of corroboration of state assessment results is a very high priority, the panel acknowledged that this is a policy activity, not a research activity. With regard to validation, however, the NVS panel urged that all components of the NAEP design be considered in determining the limits on validity of inferences that policymakers might wish to make on the basis of NAEP results.

Also considered unquestionably "Essential" are studies to compare the alignment of NAEP with state assessment standards and instruments. Within the context of P.L. 107-110, there is great concern that people who disagree with NAEP results in a particular state will argue that NAEP is not assessing skills in the form that the state mandates. There may be 50 different sets of standards for reading and mathematics, and there are definitely differences between NAEP and each state assessment. The panel felt that it may well fall upon NAEP to measure the alignment of its standards with each state's standards, leading potentially to 50 expert panel studies in reading and in mathematics, and that NAEP should prepare to carry out these studies.

Finally, the panel was unanimous in assigning an "Essential" rating to bridge studies to maintain trend reports in the context of changes in measurement. Each year new ideas emerge for enhancing NAEP's sampling, measurement, analysis, and reporting. It is also critical for NAEP to keep pace with changing concepts of educational achievement by periodically updating its frameworks in each content domain. For all these reasons, change is necessary, and although attempts to measure trends when the yardstick changes must be approximations, the approximations can be dramatically improved by conducting bridge studies to measure the effects of the yardstick change. Typically, these studies present the same participants with alternative (old and new) forms or compare results of old and new analyses of the same data.

High to Essential Studies

Two additional study areas were considered essential by more than half the panel members and had average ratings that placed them between "High" and "Essential": research to develop valid scoring of accommodated test performance and research to develop methods for adding domain item sampling error to the estimation of measurement error.

A wide variety of accommodations are in use in state assessments, and these are being introduced gradually into NAEP. Some of the more common accommodations are increased time, small-group administration, alternative language forms, and the use of a "scribe" to record responses for students. These accommodations, or non-standard test administrations, make the testing situation feasible for students with disabilities or limited English proficiency by removing specific barriers to the demonstration of their achievement, but they may also make the test easier by removing parts of the target skill domain—skills that students in non-accommodated conditions must also master. Although research to address this validity question is complex, extensive, and expensive, most members of the NVS panel considered it "Essential" that NAEP carry out studies to determine how to score accommodated performance so that groups of accommodated and non-accommodated students with the same level of content proficiency obtain the same distribution of NAEP scores.[1]

Finally, the NVS panel considered that NAEP must not be found to be underestimating measurement error and that the component of error owing to sampling items from a domain must be estimated. In fact, the development of a standard method for including this error component in overall standard error estimation is under way, and the panel considered that to be prepared to respond to challenges about NAEP's statements of precision and statistical significance, the need for this work is at least "High" and probably "Essential."

Highly Important Studies

Seven problem areas were considered sufficiently likely to raise serious threats to NAEP's validity to warrant a "High" need for studies to address them. Three of these involve gaining a more thorough understanding of the constructs being assessed by NAEP and the processes that students go through in responding to items, especially identifying irrelevant skill requirements that may contaminate test scores.

---

[1] These ratings were counterbalanced by one panel member who rated this area of inquiry as "Not Necessary" for NAEP.

The NVS panel generally felt that NAEP needs to be able to respond to challenges that contend that the items are not testing the skills described in the framework. For example, NAEP requires more writing in the reading assessment than do many tests currently in use, which may mean that students who can read very well but have difficulty writing will not perform as well on the reading assessment as students who are less proficient readers but more facile writers. If this effect is substantial, then NAEP might tend to corroborate gains more of reading programs that simultaneously build writing skills than of programs that focus on other aspects of reading. The list of ways that test items can be "contaminated" by unintended skill requirements is long, and the necessary validity studies will have many research questions.

Two other study areas rated at "Highly" important involve 1) monitoring sampling and the methods of representing the student population and 2) developing methods to represent excluded students. If NAEP is to corroborate state assessment reports of gains at the aggregate level, then NAEP is open to challenge if its sample figures cannot be weighted appropriately to represent the same population of students that the state's own program assesses. Sources of potential bias include lack of completeness in the lists from which schools are sampled or students are sampled within schools. Lack of participation by either schools or students can also lead to bias. In practice, these population discrepancies may be small, and they may have an even smaller impact on aggregate results if under-represented students tend to perform at approximately the same levels as other students. In the panel's judgment, however, NAEP should monitor each source of potential population bias and make preparations so that studies to correct for sampling and participation problems can be implemented quickly when signals are received that they are necessary.

With regard to students excluded because of disabilities or limited English proficiency, variations in exclusion rates over time have already led to challenges of the validity of NAEP's reports of gains in some states. Panel members therefore attached "High" importance to research into methods to include "excluded" students in population estimates.

Finally, the other two "Highly" needed areas involve validating changes in analytic methods that can shorten the time required for analysis and assessing the gains in trend analysis that can be obtained by using data from more than two points in time.

The new deadlines for reporting primary NAEP results will be substantially shorter than NAEP has been able to meet in the past, and shortcuts in analysis may be the only way to meet these deadlines. However, the shortcuts must be shown to produce results that are both valid and consistent with the more extensive analyses that NAEP has employed in the past. The major innovation considered is the production of direct estimates, not involving imputed plausible values and not relying on a broad range of contextual information (i.e., "conditioning") to increase the precision of population estimates.

With regard to trend analysis, analytical techniques that consider trends over more than two points in time are potentially of great use when, as is often the case in NAEP, the "signal," or expected change, is small in relation to the amount of noise in the data.

Research into trend analysis was the last of the study areas to be rated at least "Highly" important.

Other Studies

Of the remaining three studies, one, evaluation of (public) audience interpretations of reports, was considered to be "Moderately" important, but the other two may be problematic. A comparison of NAEP content with what is taught in the classroom must be carefully constrained so that it does not promote a national curriculum standard to the detriment of individual state curriculum standards. And developing methods to support certain secondary analyses while suppressing others as misleading, although potentially raising the quality of statistical policy analyses, can also raise issues about freedom of access to government-produced information.

Summary

Thus, the recommended NAEP validity research agenda consists of research in four essential areas, nine highly needed areas, and three less important areas. In this phase of setting a validity research agenda, the NVS panel did not consider either the cost of the studies or specific design issues. Some of the studies requiring new data are likely to be the most expensive and to require the greatest time; studies based on expert panel judgments or cognitive lab studies (which gather new information, but not in the same amounts as the "new data" studies) are less expensive and time-consuming, and analyses that use existing data in new ways are generally the least expensive and the least time-consuming. Nevertheless, there is a wide variation within each category (e.g., how many states will be included in separate alignment studies?).

Finally, we should note what is not included in this set of recommended studies. The NVS panel did not address the issue of conscious cheating, which would certainly invalidate results, primarily because that is an "auditing" rather than a "research" function. And the NVS panel did not focus on issues of developing new frameworks and new topic areas for NAEP, such as testing students' abilities to use computers for writing or their abilities to work as team members. Instead, the panel focused on threats to the validity of NAEP reports in the current context, extended to include the imminent use of NAEP as a check on independent state assessment results.

Chapter 1. Introduction

NAEP Background

At the beginning of the 21st century, the National Assessment of Educational Progress (NAEP) continues to stand as a unique, and uniquely valuable resource for American education. It is the only periodic measure of student achievement based on national probability samples, and it is the only method by which states can validly compare the academic progress of their students against common high standards. For over 30 years, NAEP has been reporting national achievement trends in mathematics, reading, and science; and for more than 10 years, NAEP has been reporting on achievement trends on a state-by-state basis, through the voluntary state NAEP.

While serving as "The Nation's Report Card," NAEP has addressed a wide range of issues in testing methodology and has maintained a level of rigor that serves as a standard for other assessment programs. NAEP constantly balances competing needs a) to report precise estimates of achievement and achievement gaps, b) to minimize testing burden, and c) to report results soon after testing. Each year, NAEP faces new psychometric issues as it attempts to provide useful information to various audiences while responding to the changing educational and social contexts of testing. These issues range from changing conceptions of what should be tested in areas like reading and mathematics to changing priorities and constraints for testing children with disabilities or limited English proficiency. Continuous vigilance is necessary to ensure that NAEP results remain valid in the face of threats to its validity.

At various points, The National Assessment Governing Board (NAGB), the NAEP Design and Analysis Committee (DAC), the National Academy of Education's Panel on the Evaluation of the Trial State Assessment (TSA Panel), NAEP's Technical Review Panel (TRP), and the National Research Council (NRC) have all watched carefully over NAEP to ensure that new demands do not compromise the validity of NAEP reports. The NAEP Validity Studies (NVS) Panel has joined in that role since 1996, examining a variety of issues surrounding NAEP development and implementation and carrying out studies to address these issues. In this report, the NVS Panel addresses NAEP's need for a comprehensive agenda for validity research.

The NAEP Validity Studies Panel

Because the issues that threaten NAEP validity are diverse, the panel to protect NAEP's validity must be interdisciplinary, including experts with a variety of specialties. The NVS Panel, chaired by George Bohrnstedt (American Institutes for Research), consists of 15 individuals with expertise in:

- Educational research—to help make sure that NAEP is meaningful
- Psychometrics—to help make sure that NAEP is accurate
- Curriculum—to help make sure that NAEP instruments are relevant
- Sampling—to help make sure that the NAEP sample represents all students
- Test fairness—to help make sure that NAEP represents subpopulation achievement accurately
- State assessments—to help make sure that NAEP addresses state needs
- Long-term familiarity with NAEP

While each member of the NVS Panel has expertise in more than one of these areas, discussions of NAEP validity are particularly informed by the presence of specialists. Leading educational researchers on the panel include David Grissmer (RAND), Larry Hedges (University of Chicago), and Lorrie Shepard (University of Colorado). The panel's psychometricians include Al Beaton (Boston College), Darrell Bock (University of Chicago), and Don McLaughlin (American Institutes for Research). Audrey Champagne (SUNY, Albany) and David Pearson (University of California, Berkeley) contribute special expertise in curriculum content, and James Chromy (RTI) provides expertise in sample design. Richard Duran (University of California, Santa Barbara) and Gerunda Hughes (Howard University) keep issues of equity for subpopulations in focus; and Gerald DeMauro (New York) and Peter Behuniak (Connecticut) contribute "reality checks" from their experiences in state assessment programs. Finally, Ina Mullis (Boston College) and Al Beaton inform considerations of "new" validity problems from their extensive experiences in the implementation of NAEP.

Validity Research

Validity is the extent to which the messages in NAEP reports accurately communicate the state of educational progress in America to educators, policymakers, and the public. If NAEP reports that the gap in student achievement in mathematics or reading is widening, many questions can be raised about the meaning of that report. Is the mathematics or reading tested by NAEP the kind of mathematics or reading skills that are important for students to achieve? Is the way that NAEP tests these skills fair and accurate? Do the results represent the full student population? Do the analytical procedures accurately tie the data to general statements about achievement? Are the results stated sufficiently clearly and unambiguously in the report that misinterpretations are avoided? What special information does valid measurement of trends require?

If the answer to any of these questions is "No," then NAEP's report of a widening gap can be challenged. Validity research consists of studies undertaken to address these questions, to prepare responses to the challenges. Opinions may differ, for example, as to which mathematics or reading skills are most important, but research studies can explain what component skills NAEP assesses. Other research can determine the extent to which aspects of the testing situation color the measurement of the assessment's target skills and the extent to which the full population is represented by the NAEP sample. Tests of the analytic procedures can verify the extent to which they support the generalizations one makes based on the outcome of those procedures (e.g., the accuracy of estimates of statistical significance).

To prepare a systematic analysis of the domain of validity threats and to identify the most urgent validity research priorities, the NVS Panel created a broad framework that encompasses all aspects of NAEP. The framework was broken down into six broad categories:

- Subject Domain—What is Being Measured?
- Subject Domain—How Is it Being Measured?
- Representing Populations
- Data Analysis
- Reporting and Use of NAEP Results
- Assessing Trends

A subcommittee of two or three of the NVS panel members assumed the responsibility for each of the six broad areas of validity concerns. Over the course of two NVS panel meetings, the subcommittee (or one key author from the subcommittee) prepared a paper on each study area, and the full panel discussed the six papers. The panel members then identified and prioritized key validity issues requiring studies by the NAEP program. The third panel meeting in November 2001 focused on reaching a consensus on priorities of different areas of validity research across the six broad categories.

These six papers, which describe the NVS Panel's recommendations for important validity studies, are presented in this report. Chapter 2, *What is Being Measured?* discusses alignment studies that address the extent of the alignment of the NAEP assessment frameworks with state and local standards, curriculum frameworks, and test frameworks, as well as cognitive requirement studies that would focus on the question of how well NAEP measures students' understanding of domain knowledge valued by the public. Chapter 3, *How is It Being Measured?* focuses on three issues: 1) the sensitivity of a comprehensive assessment to instructional variation, 2) issues of test bias, and 3) effects of teaching the test. The proposed study would gather criterion performance data from multiple sources to evaluate and verify whether various components of NAEP provide a "true picture" of student proficiency.

Chapter 4, *NAEP Validity Issues: Representing Populations,* proposes four categories of validity research studies to address threats to validity arising from: 1) incomplete lists of schools, 2) nonparticipation of schools, 3) incomplete lists of students, and most urgently, 4) nonparticipation by students, which includes students who fail to appear for the assessment and students who are excluded. Chapter 5, *Issues and Recommendations on NAEP Data Analysis*, outlines three procedures that may contribute to simplifying and speeding up the analysis of NAEP data: 1) multiple-group item response theory (IRT), 2) a similar MML procedure that estimates regression relationships among examinee background characteristics, and 3) item bi-factor analysis. The chapter also considers the potential advantages of using the school as the unit of analysis or including item sampling in the estimation of standard errors.

Chapter 6, *Validity and Utility Issues in NAEP Reporting and Data Releases*, proposes that a priority for NAEP validity research must be a focus on the validity of NAEP's conclusions regarding states' progress. This research would broaden the definition of "reporting," and include considerations of how NAEP findings are presented to the public, and who is to decide whether NAEP results corroborate state progress. Chapter 7, *Estimating Trends from NAEP Scores*, explores whether more sophisticated multi-point trend analysis methods could improve the validity of measures of progress and outlines a research agenda that determines when two-way comparisons are appropriate, and when trends are appropriate.

Finally, Chapter 8, *Synthesis: An Agenda for NAEP Validity Research*, is a summary resulting from an extended panel discussion of the relative importance of each of the studies, or areas of study, proposed in the earlier chapters. Twenty-two studies, or areas of study, were identified and each panel member gave a priority rating to each study or area of study. Panel members then identified "essential" study areas that are needed to address threats to the validity of NAEP reports.

Chapter 2. Subject Domain: What Is Being Measured?

Subcommittee: Audrey Champagne
P. David Pearson

Rationale for Studies of the Subject Domain

Critiques by educators, subject domain experts, and representatives of the academic disciplines inevitably follow the release of NAEP student performance data. Often the critiques derive from the released items that are used to inform the public about the subject matter on which the performance data are based. Both the form and content of the released items are the subject of criticism.

Teachers and school-based subject matter coordinators criticize large-scale mathematics and science tests, claiming that their students understand the mathematics and science contained in released items, but that the reading demands of the items are so great that they cannot perform well despite their understanding of the content domain. Teachers and coordinators level similar criticisms against items requiring extended responses. Teachers and coordinators claim that the students understand these items but cannot express that understanding in written form. Ironically then, two other domains of NAEP assessment, reading and writing, may be interfering with our capacity to assess mathematics and science with high degrees of validity.

Representatives of teacher and coordinator professional societies, argue that many of our tests, including not only mathematics and science but also reading and writing assessments, do not represent the subject matter content valued by educators. This is especially true of wide-scale tests used for accountability purposes. Often the criticisms focus on what these critics claim is the over-representation of items measuring lower level information and the under-representation of items measuring higher level cognitive abilities, such as problem solving or inquiry.

These same educators are critical of the alignment of content on NAEP with students' opportunity to learn, claiming that the subject domain sampled by NAEP assessments does not correspond with the requirements of state standards or the frameworks from which state mandated tests are developed.

Representatives of the academic disciplines criticize the choice of principles tested, claiming that they do not represent the most powerful and newest ideas of the discipline. Discipline-based critics claim that multiple-choice items do not assess true understanding. They are also highly critical of the accuracy of the items, pointing out, for example, that a response scored as correct may not be *exactly* correct in the context described in the stem of the item.

The advent of President George W. Bush's proposal to use NAEP as the criterion for evaluating the validity of state tests that monitor mathematics and reading performance raises a significant policy issue related to the subject matter domain. If the subject matter domains sampled by NAEP are different from state mandated content for those same domains, the states can claim federal infringement on their right to control education.

NAEP must be prepared to address the criticisms of the stakeholders. Focusing on the subject domain issues places certain relevant social and philosophical criticisms in the background. However, the technical issues relevant to the subject domain are highly complex, even ignoring, for the moment, the consideration of the social and philosophical issues underlying these debates and dilemmas.

The criticisms cluster about two central nodes: 1) alignment, and 2) the cognitive requirements for successful performance. Ultimately, the alignment issues revolve around the question of how well the curricula students have experienced aligns with the performance expectations of NAEP. The cognitive requirements for successful performance revolve around the question of how well NAEP measures students' understanding of domain knowledge valued by the public (or for that matter, anybody claiming some authority over curriculum).

Alignment

Alignment is difficult to study. Questions involve many different forms of documentation (standards, tests, test frameworks, curriculum descriptions) that exist in different versions at different levels—national, state, and local. Alignment also involves the relationships among a) the documentation that represents what it is *intended* for students to learn, b) the ways in which individual teachers *implement* the intentions, and c) how the implementation interacts with student populations to result in *learning*. In shorthand terms, this is the distinction among what is intended, what is implemented, and what is learned.

The inclusive study of alignment requires assessment of the degree of congruence among standards, tests, test frameworks, and curriculum descriptions within a jurisdiction (national, state, local), as well as the degree of congruence of each across jurisdictions. The congruence between conceptual frameworks and translation of the frameworks into action (implementation) is also a part of the inclusive study of alignment.

Figure 2.1–Alignment Across Documents

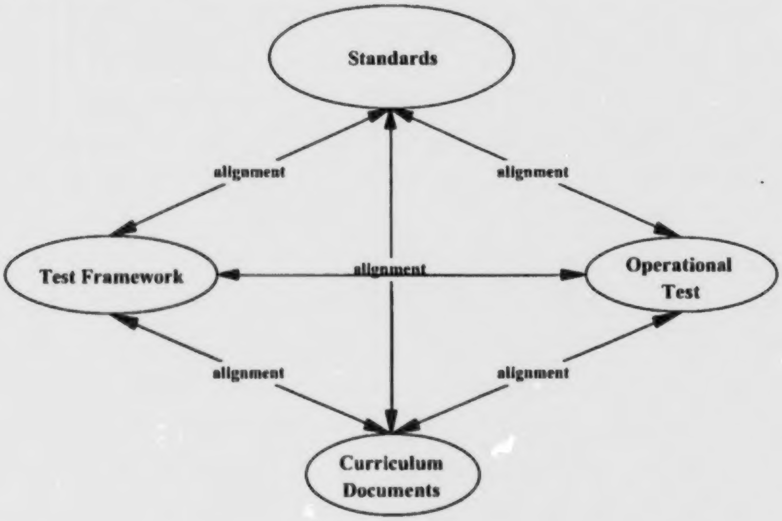
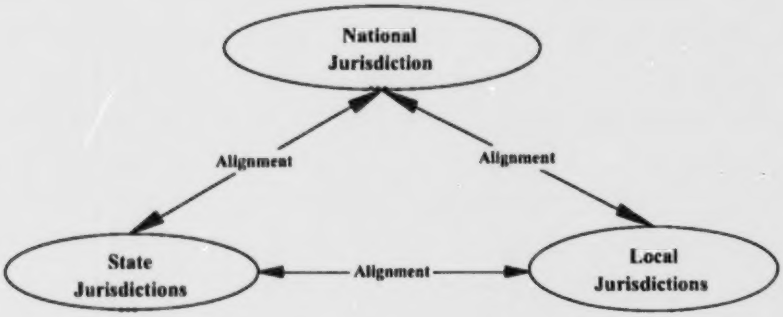

Standards
alignment
alignment
Test Framework
alignment
Operational Test
alignment
alignment
Curriculum Documents

Figure 2.2–Alignment Across Jurisdictions


National Jurisdiction
Alignment
Alignment
State Jurisdictions
Alignment
Local Jurisdictions

Not only is the inclusive study of the many facets of alignment inherently difficult, it is also challenged by the fact that the language describing the subject matter domains and the levels of performance expectations is neither well established, nor consistent from one jurisdiction to another. There is little agreement within the many communities with interest in, and responsibility for, assessment about the meaning of some very basic concepts from the natural sciences, education, or psychology. *Density* is a good example because it is ubiquitous in educational, disciplinary, and testing settings in all of the natural sciences. Its characteristics are not unusual. *Equivalence* is an example from mathematics that has similar characteristics. *Higher-level comprehension* in reading exhibits the same set of problems within reading assessment.

To illustrate these issues more concretely, we have chosen to unpack them around the *density* construct. They could just as easily have been illustrated for *equivalence* or *higher-level comprehension.*

What does it mean to understand density? Standards call for students to understand density. Curriculum documents provide activities and strategies for developing that understanding. Density appears as a topic in test frameworks, and items testing the understanding of density appear on tests.

Should eighth graders who have had the opportunity to learn about density be able to

- Manipulate the formula, $\sigma=m/v$, calculating the density, mass, or volume given the two other quantities
- Calculate the density of a regularly shaped object such as a cube or cylinder given its mass and linear dimensions
- Calculate the density of an irregularly-shaped object such as a plastic tea cup
- Describe how to measure the density of an irregularly shaped object such as a stone
- Accurately measure the density of an irregularly-shaped object such as a stone
- Predict how an object that floats with ¼ of its volume under pure water will float in ocean water or salad oil
- Explain why a steel ship the size of a football field floats, while a cube of steel 1 centimeter on a side sinks
- Calculate the buoyant force on a ball floating on the surface of a pond
- Calculate the net force on an air filled ball held 3 meters under water
- Explain the motion of the ball at the instant it is released
- Explain why Archimedes' bathtub overflowed when he got in it, and explain how Archimedes' observation helped him solve the problem of how to decide if the king's ring was pure gold

Neither standards nor a test framework stating that eighth graders should understand the concept of density and be able to apply it, define what it means to understand and apply the concept. Only items that require students to perform tasks such as those described above make explicit what it means to understand and apply the concept of density.

Cognitive Demands

The complication of a lack of common understanding of subject domain concepts is closely related to the second cluster of criticisms, those surrounding the cognitive requirements for successful performance. Based on expert judgment, items are categorized as requiring recall of factual information, understanding, application, problem solving, or some other descriptors. Alignment of a test with the test framework is based on just such expert judgment. In turn, the framework is based on expert judgment of the appropriate sample of the subject domain to be tested. What is not known is how well expert judgments about knowledge and cognitive processes underlying recall, understanding, application and problem solving match the knowledge and cognitive processes that students actually apply in responding to the items. In other words, do the test items actually elicit the sorts of cognitive processes or tap the subject matter knowledge intended?

The issues surrounding the cognitive requirements of an item is further conflated by the way in which the item is presented to the student and how the student is required to respond. This difference between what is being measured and how it is being measured is difficult to sort out.

Is being able to read a "fair" requirement for a science or mathematics test? Is being able to read or write about science or mathematics a "fair" requirement for a reading test? Is being able to interpret information presented in graphic or diagrammatic form a "fair" requirement for a science test, or, for that matter, a reading test? Are responses in written, graphic, diagrammatic, or demonstration forms (for instance, using a meter stick to measure length) "fair" requirements for science or mathematics tests? Incidentally, the writing requirement applies equally as aptly to reading: To what degree does writing compromise our capacity to make unambiguous judgments about reading processes? Ultimately, the answer to the fairness question is whether or not reading, writing, and communication in other representations are defined as a requirement of the test. Empirically, format and representation raise two important questions:

- When items are communicated to students in alternative representations, do they measure the same subject matter domain knowledge or understanding?
- Do alternative response formats measure different subject matter domain knowledge or understanding?

Yet another complication arises from the proposal to test higher-level abilities, such as problem solving in mathematics, inquiry in science, or text interpretation in literature, in group settings rather than exclusively in individual settings. The use of group settings, usually raised in the name of greater ecological validity for the assessment, raises interesting issues around what we are measuring and why we are measuring it. With regard to the *what*, successful engagement in group activities requires more than knowledge and abilities related to science, mathematics, or reading. Being socially astute, knowing when to lead and when to follow, and being sensitive to the needs of group members are social skills, not reading, mathematics, or science content. Is there an assumption that having students inquire or problem solve in groups will provide more information about their reading ability and science and mathematics problem solving or inquiry skills? *Or*,is the purpose to learn about social skills? *Or* is the purpose, as we have implied, to simply increase the ecological validity of the assessment? Success in the business world and active engagement in the democratic process may require social skills. However, that does not mean necessarily that social skills should be measured in a reading, science, or mathematics test.

These questions raise another more daunting policy issue—whether NAEP should shape testing and curriculum at the state and local level. Demonstrating the nature and intensity of factors influencing student achievement will impact policy at the national and state levels. Consequently, studies aimed at assessing such influences are important if NAEP is to serve its policy functions. For example, studies documenting in finer detail the cognitive requirements for performance have considerable potential for influencing classroom practice. Similarly, studies that help teachers provide students with opportunities to develop subject domain understanding will serve NAEP's policy functions.

Proposed Studies

Alignment studies might be designed to address the research question: What is the extent of the alignment of the NAEP assessment frameworks with other documents that purport to provide valid and adequate representations of a subject matter domain—national standards, state standards, state or local curriculum frameworks, test frameworks, or tests themselves? Evidence of congruence can and perhaps should be used to evaluate the validity of claims regarding the relevance of NAEP to states' educational responsibilities.

Partial answers to this question can be obtained with alignment studies that assess the congruence between NAEP frameworks and state standards, curriculum frameworks, or test frameworks. The choice of which state documents to select depends, at least in part, on the availability of the documents. If all of the documents in a given state are aligned, it will not make any difference which document is selected. However, the assumption of alignment among state documents is tenuous and probably needs to be tested empirically. The best choice of state document is probably the state's mandated test framework. The choice is reasonable because test frameworks have the same function and similar forms across jurisdictions, and because classroom teachers often teach to a high stakes or highly visible test. (The second reason is valid only if the state mandated test is aligned with the test framework.)

The alignment study requires a prescriptive design of the process for measuring alignment, and specifications of the characteristics of the individuals who would constitute the expert panel charged with carrying out the measurements. Selection of the sample of states (about 12) to be a part of the study would be based on criteria such as 1) performance of the state's students on the particular NAEP assessment under study and 2) the degree to which the state controls local curriculum. It also would be useful to sample some high profile assessment policy states (e.g., Texas, California, Massachusetts, or Kentucky) and some quieter (from the policy perspective) states (e.g., Iowa, New York, or Wyoming).

The second level of alignment studies, which would index the degree to which various assessments, be they NAEP or state assessments, influence opportunity to learn in our schools, are both more important and less feasible. They are important because we need to know the degree to which 1) tests influence opportunity to learn in any given domain, and 2) the efficacy of any such influence on students' acquisition of knowledge and skill in that domain. They are less feasible because of the enormous cost in creating a chain of connection, not to mention causality, between NAEP documents, state documents, and classroom implementation of curriculum.

The design of the cognitive requirements studies is much more complex depending upon whether a particular study will focus on 1) the influence of reading and writing (or other forms of verbal or representational communication) on performance in discipline-based school subjects (mathematics and science), or on 2) the knowledge and cognitive processes that students actually apply to their solutions of NAEP items. At the very least, we should capitalize on what we have learned in our own studies of the cognitive demands of reading and mathematics tasks[1] and in the cognitive laboratory work carried out by the American Institutes for Research.[2] What would be ideal is a combination of cognitive laboratory studies, in which we watch and listen to students as they take items from NAEP assessments, and case studies of schools and classrooms, in which we try to link cognitive laboratory performance and perceptions about items with knowledge about teaching and learning in particular instructional settings. If we were able to gain access to a few sites with known curricular characteristics, we would be able to learn much more about both the cognitive requirements of different NAEP tasks and the relationship between performance on various task formats and opportunity to learn.

---

[1] Pearson P.D. & Garavaglia, D.R. (1998, April). *Improving information value of constructed response items when mixed with multiple-choice items.* San Diego, CA: American Educational Research Association.

Either line of research, alignment or cognitive requirements has potential for greatly increasing the influence of NAEP in the reform movement. Both are costly and complex, but that should not decrease our conviction to carry out the work.

---

[2] Pane, N., & Levine, R. (2001). *Cognitive labs: An essential test development step.* 31st Annual Conference on Large Scale Assessment, sponsored by the Council of Chief State School Officers, Houston, TX.

Levine, R. (1999). *New item development technologies: Cognitive labs.* 29th Annual Conference on Large Scale Assessment, sponsored by the Council of Chief State School Officers, Snowbird, UT.

Paulsen, C., & Levine, R. (1999). *The applicability of the cognitive laboratory method to the development of achievement items.* Presentation at the American Educational Research Association, Montreal, Canada.

Chapter 3. Subject Domain: How Is It Being Measured?

Subcommittee: *Lorrie Shepard*
Richard P. Duran
Gerunda Hughes

For almost two decades, concerns about the performance of American children on national and international assessments—initially prompted by *A Nation at Risk* (1983)[1] and later by *Goals 2000: Educate America Act* (1989)[2] and the Third International Mathematics and Science Study (TIMSS)—have compelled us to extend our thinking beyond what used to be traditional views of curriculum, instruction, and assessment. Years ago it was assumed that a mathematics curriculum focused primarily on numerical manipulations and calculations, that instruction was primarily didactic and teacher-centered, and that items on achievement tests generally required students to reproduce information as it was presented in class. At the same time, it was assumed that an achievement test had content validity if the items on the test were judged by content experts to measure the right content in the right proportions.

Presently, the boundary lines of the subject domains are blurred. Subjects are no longer taught in isolation. Approaches to teaching and learning require more active engagement on the part of the learner. And much of what and how we measure achievement in some content areas reflects new principles and standards of learning. For example, writing in mathematics reflects the standard of communication, and integrating mathematics with other content areas to solve real world problems reflects the standards of problem solving and connections. For sure, the "it" in the above title has changed over the years. Correspondingly, how "it" is being measured has changed, and these changes have serious implications for addressing the full range of issues related to test validity.

Validity Issues

Just as curriculum standards and expectations for what students should know and be able to do have become more challenging over time, so too have the requirements for evaluating test validity. Instead of merely confirming that a test includes the right content, more contemporary validity standards require empirical verification that a test actually measures subject area proficiency as intended. We need to separate questions about the construct being measured from questions about whether the test problem embodies those constructs. In particular, it is important to verify that students are not prevented from demonstrating their knowledge and skills because of artifacts of the assessment format. Nor should they be allowed to boost their scores artificially by extended practice or familiarity with the test format.

---

[1] *A Nation at Risk.* (1983). Washington, DC: United States Department of Education.
[2] *Goals 2000: Educate America Act.* United States Congress. Washington, DC: Author.

For example, it can be argued that when we construct tests that require students to write as a means to demonstrate their reasoning, or when we administer tests via computers, extended practice and familiarity with the format becomes an important part of being able to demonstrate what the student knows and can do. On some tests or assessments, the format is so inextricably tied to what is being assessed that performance becomes a function of both "what is being measured" and "how it is being measured." In essence, "artifacts" under older ways of assessing have become "expectations" under newer curriculum standards. Nonetheless, to merely assume that all students have been exposed to more challenging curricula and contemporary problem contexts could put students in jeopardy of having their knowledge and skills underestimated. Therefore, it is important that the effects of various assessment formats and demand characteristics be directly investigated.

Although past research cannot tell us automatically about the validity of new tests or new test uses, past research can alert us to the most likely threats to validity and help target future research investigations. Here we consider three issues of particular importance to the validity of NAEP:

1. The sensitivity of a comprehensive assessment to instructional variation

2. Issues of test bias

3. Effects of teaching the test

Comprehensive Assessment and Instructional Variation

"Alignment" has become the watchword for assessments developed as part of state accountability systems. For accountability purposes, standards, curricula, assessments, instructional materials, and professional development should all be aligned so that they can be used in concert to improve student achievement. In contrast, it is important to note that NAEP was not designed as part of an accountability system. Rather, NAEP was intended to serve as an independent monitor of educational progress, more like the U.S. Census or economic indicators. In keeping with its purpose as an indicator, the content of NAEP must be "comprehensive".[3] It cannot be aligned with any one version of content standards but must be inclusive of curriculum standards across the 50 states and across time. A comprehensive assessment domain that includes both traditional and forward-looking content is needed to detect the effect of policy changes on student performance. During the 1980s, for example, the breadth of content sampled in NAEP made it possible to document a decrease in higher-order thinking skills at the same time that performance on basic skills increased. If NAEP content had been narrowed to focus only on basic skills policies popular at that time, the assessment would not have been able to capture these diverging trends.

---

[3] Glaser, R. & Linn, R. (Eds.). (1992). ***Assessing student achievement in the states: The first report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment.*** Stanford, CA: National Academy of Education.

NAEP is a large-scale survey rather than a tightly controlled research study; therefore, NAEP results cannot be used directly to establish the cause of improvements or decrements in student performance. Nonetheless, NAEP should be able to detect and report on significant shifts in student performance associated with major policy decisions. We refer to this as the "instructional sensitivity" of an assessment. The study proposed here samples a range of instructional contexts to make sure that students are able to show what they know, regardless of context. In other words, if the assessment format should not favor one type of instruction or curriculum over another, then the match or correlation between test results and criterion performance should be the same across contexts. In addition, the comparisons of results from different contexts should also establish the sensitivity of the assessment to instructional variation. If the assessment can capture these kinds of effects in cross-sectional data, then there is a good likelihood that it can be used to monitor such differences longitudinally as well.

Test Bias

The literature on test bias is concerned with the many instances when students "really do know" a concept, but are prevented from showing what they know by some unnecessary, construct-irrelevant difficulty in the test. This may be the result of an aspect of item formatting or of the mapping from the construct onto the physical presentation of an item. The most pronounced case is that of English language learners for whom all tests become a measure of English reading proficiency regardless of whether the test was intended to measure knowledge in mathematics, history, or science. Although large-scale assessments use accommodations in an attempt to address the more extreme instances of biased measurement, many more subtle forms of distortion or invalidity affecting some groups and not others persist. For example, extensive research on Advanced Placement Examinations shows that females do relatively better on essay tests, while males do better on multiple-choice questions.[4] Determining whether either of these tilts in test format—favoring one group or the other—is evidence of bias would require further evaluation of the test construct and additional evidence of criterion performance, i.e., who really knows what. In the proposed study, we would gather criterion performance data from multiple sources as a means to evaluate and verify whether various components of NAEP give a true picture of student proficiency.

Teaching-the-Test Effects

In contrast to the research on test bias, where students are hindered from showing what they know, the research on teaching the test warns us of the reverse problem, where students can appear to know content they have not really mastered. Narrow teaching to the test can produce inflated test score gains, meaning that test scores go up without there being a generalized increase in knowledge. For example, Koretz, Linn, Dunbar, and Shepard conducted an experimental study to evaluate whether reported test score gains were "real."[5] In large school districts, selected because of their high-stakes pressure on scores, random subsamples of students were administered unfamiliar standardized tests and alternative tests that were constructed item-by-item to match the district-administered test, but using a slightly more open-ended format. Student performance dropped as much as a half standard deviation on the unfamiliar tests, suggesting that students did not really know all that they appeared to know on the publicly reported measures.

---

[4] Schmitt, A.P., Mazzeo, J., & Bleinstein, C. (1991, April). *Are gender differences between Advanced Placement multiple-choice and constructed response sections a function of multiple-choice DIF?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

In the same study design where classroom samples of student work and state tests would be used as a check on the validity of NAEP, NAEP could be used to check on the generalizability of results from the state test. To escape the circularity of using each test as a check on the other, individual assessments would also be used in combination with samples of student work to establish a more certain representation of student proficiency. Then any discrepancies between NAEP, the state test, and verified level of proficiency (classroom work + individual assessment) could be analyzed in conjunction with student and teacher data on instructional practices to determine whether differences in results were due to true differences in curricular goals, test bias, or teaching-the-test effects.

Study Methods

Selection of States and Schools

As described below, the study calls for the collection of multiple measures of achievement for a sample of students in addition to their NAEP "scores."[6] Ideally such a study would be coordinated with field trials or the operational administration of NAEP, and be conducted immediately following the group administration. If such coordination were not possible, then the study would also entail concurrent administration of NAEP booklets.

Three to four states should be selected to reflect important differences on two dimensions: 1) the amount of teaching the test to be expected from high-stakes accountability pressure, and 2) the amount of emphasis in the state curriculum and assessment on basic-skills versus more advanced content standards.

Within each state, 18 classrooms should be selected to represent high-performing, middle-performing, and low-performing schools as identified by state test results (six schools in each of three strata). The 18 classrooms should come from 18 different schools. Assuming 25 students per classroom (each with data available on NAEP and the state test), the sample size in each state would be 450 students.

---

[5] Koretz, D., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

[6] NAEP does not typically produce scores for individual students, but for purposes of the study scores could be estimated using NAEP booklets or pairs of booklets.

Individual Assessments, Classroom Data, and State Assessments as Criterion Measures

A set of validity criteria would be established against which the validity of the group-administered, large-scale NAEP test can be evaluated. The criterion measures would include individual assessments, classroom samples of student work, and state assessment data. Individual assessments would be developed and administered as a central component of the study. These individual assessments would cover more of the intended and taught curricula content than is represented on NAEP. Intended content would be ascertained from the NAEP frameworks as well as curriculum framework documents for the four states in the study.

The individual assessments would be used to evaluate whether the usual NAEP administration gives a true picture of what students are taught and what they know and are able to do, both in relative and absolute terms. Furthermore, if the individual assessments are administered both orally and in writing (usual format), additional evidence may be obtained about student learning. Content coverage on the official NAEP and individual assessments should be examined for overlap and alignment, and discrepancies between the official NAEP and individual assessment results should be examined for patterns and for explanations as to the cause of those patterns. Are discrepancies greater for certain groups of students? For certain types of assessment tasks? Or in certain types of instructional settings?

Classroom data would take two forms. First, teachers could provide samples of the assessments that they use in their classrooms. Examination of these artifacts would help determine whether, in the day-to-day learning environment, the form and content of their classroom assessments reflect the new principles and standards of thinking and problem solving in the subject domains, and the new ways in which knowledge and skills in the subject domains are measured on state and national assessments.[7] Second, both teachers and students should be asked to provide additional sources of validity evidence. Teachers would be asked to provide

- Relative ratings of students within each class
- Identification of NAEP tasks that have been taught extensively, taught to a limited extent, and not taught as part of the local curriculum
- Judgments of individual student proficiencies on specific tasks that anchor the NAEP achievement scale
- Examples of students' class work relevant to the achievement continuum
- Possible explanations for any large discrepancies observed between group NAEP results and individual assessment results

Students should be asked

- What material was taught?
- Did they understand the material when it was taught? To what extent?
- How much time do they spend studying outside of class?

---

[7] Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29* (7), 4–14.

- What perceptions do they have about their teacher's expectation of their performance?

State assessment data should also be gathered for the students in the study sample to serve as an additional validity criterion. In addition, any systematic differences between state and NAEP results should be analyzed. Are any relative strengths or weaknesses on NAEP associated with corresponding emphases or omissions in the state assessments? For example, students in a state with only a basic skills state assessment may do relatively better on basic skills items on NAEP, but score relatively less well on problem solving and items requiring higher-order thinking skills. If this were the case, we could be more assured of NAEP's validity to document the effects of differences in instruction.

Ideally, examiners on site would conduct some amount of data integration so that the study would provide more than a series of correlation coefficients, which are difficult to interpret. For example, examiners might try to reconcile the several sources of evidence on an individual student. Do teacher ratings, classroom work, official NAEP, the state assessment, and the individual assessment all tell the same story of student proficiency? When there are discrepancies, can triangulation and weighing of evidence be used to decide which results are more credible? While on site, are there a few students with the greatest discrepancies who could be observed further? What do these combined criteria tell us about the validity of NAEP? Are NAEP results consistent with other sources of evidence taken together, or is NAEP often the outlier? Are factors such as student motivation and completion, artifacts of test format, curricular alignment, reading and writing demands for low achievers, special education students, or English language learners associated with occasions when NAEP is the outlier?

It should also be noted that some instructional strategies have been found to be more effective with some subgroups of the population than for others. For example, it has been repeatedly documented that African American students learn best in learning environments that use cooperative and communal groups.[8] Yet, these types of assessment formats are rarely used in large-scale assessments even though they may be practiced in classroom teaching and assessment. (The Maryland School Performance Assessment is an exception.) Rather than only asking how well classroom practices are aligned with NAEP, the study should also identify ways that NAEP would have to change to capture or give credit for competencies seen in the classroom that are not seen on NAEP. Both students and teachers in the individual assessment study should be surveyed to identify instructional content and even samples of student work that speak to content area competence not reflected in the NAEP assessment.

Motivation Effects

If two assessments intended to measure the same thing produce different results, we cannot be sure whether the difference is caused by true differences in content, differences in test characteristics, or differences in the context of administration. Each competing hypothesis has implications for additional studies or safeguards within the initial study design. One very likely confounding effect is the likelihood that students will stay on task and work longer in an individual assessment than during the official, group-administered assessment. As part of the individual assessment study, teachers should be asked to describe the conditions of administration and any prior exhortations received or given to ensure maximum performance. Data should be examined to see if completion rates are greater when NAEP is treated as high-stakes by local administrators and whether there are, correspondingly, greater correlations between state assessment results and NAEP under such circumstances. Students should also be asked what they were told about the importance of the assessment, and asked about how hard they tried on the official assessment.

---

[8] Boykin, A.W. (1996, April). *A talent development approach to school reform: An introduction to CRESPAR.* Paper presented at the Annual Meeting of the American Education Research Association, New York.

Irvine, J.J. (1990). *Black students and school failure.* Westport, CT: Greenwood.

Ladson-Billings, G. (1994). *The dreamkeepers: Successful teachers of African American children.* San Francisco, CA: Jossey Bass Publishers.

Madsen, J.A. & Mabokela, R.O. (2000). Organizational culture and its impact on African American teachers. *American Educational Research Journal. 37* (4). 849–976.

To isolate the effects of individual administration alone as the cause of differences in performance, a random subsample should be given individual assessments identical to NAEP.

Format and Other Test-Specific Effects

To pursue the hypothesis that students may be hindered by test format from showing what they know, the individual assessment study should include both prompted and unprompted conditions (again with random subsamples). In the prompted condition, the examiner might ask whether students have ever had to do similar problems before, and he or she could provide structured hints to make sure that students know what the question is asking.

Going further, special instruction could be provided to a random sample of students on both how to maximize correct responses to an item format and how to process the grammar, vocabulary, and visual displays used to map constructs onto specific item content. If that instruction increases scores, one is forced to ask the questions about whether that type of instruction is what we want to be happening in schools.

Special Populations

If possible, it would be desirable to oversample both special education students and English language learners to examine validity issues unique to these populations.

Chapter 4. Validity Issues: Representing Populations[1]

Subcommittee: *Donald H. McLaughlin*
Peter Behuniak
James R. Chromy

Overview

NAEP, the Nation's Report Card, publishes statistical estimates of the achievement of populations of students. In national NAEP, the populations are students in grades 4, 8, and 12 in public and private schools in the United States. In state NAEP, the populations are students in grades 4 and 8 in public schools in each participating state.[2] In addition to overall estimates of achievement, NAEP publishes similar estimates for subpopulations, such as male and female students and students with different race/ethnic identifications.

NAEP's estimates are based on the performance of samples of students, and a fundamental issue for the validity of NAEP is whether the samples of performance are being combined in a manner that ensures that the published statistics are representative of the intended populations. It is essential, for trend comparisons over time, comparisons between states, and measurement of the gaps between subpopulations, that each population group being compared be accurately represented in the estimation process.

From the statistical point of view, the issue is whether every student in the population has some known non-zero probability of being in the sample whose performance is measured. When this is true, careful processing of the data can ensure accurate representation of populations. However, if some students have no probability of being included in the sample, then those students are not represented in NAEP publications.

To select the students whose performance will yield estimates for the population, NAEP starts with a list of schools. It selects a sample of schools with known probabilities, and for each selected school that agrees to participate in NAEP it obtains a list of students in the target grade.[3] It then selects a sample of students from the list to participate in the

---

[1] This paper benefited from useful comments by Keith Rust.

[2] In the fine print, NAEP acknowledges departures from these simple population descriptions, such as the exclusion of students with disabilities who are unable to participate in the assessment; but the intent of NAEP, and the interpretations most readers place on NAEP publications, refer to these simple descriptions.

[3] In the past, NAEP has sampled by age (9, 13, and 17) rather than, or in addition to, grade. This discussion focuses on the representation of a grade-level population. Issues surrounding the definition of "fourth grade," questioning the comparability of February scores for students who start fourth grade in August or October, deserve at least preliminary examination. For interpreting state NAEP scores, where interest in year-to-year gains is intense, it is important that the school-year calendar not change significantly in a state between two NAEP administrations.

NAEP testing session. Most, but not all, of the students selected to participate provide test performance that can be used to generate population statistics.

Thus, ideally, every relevant student is in a school on NAEP's list, is on the list of students in the school that NAEP uses, and completes the assessment, so that all students are represented in the achievement statistics NAEP publishes. Validity research studies are needed that focus on the impact of departures from this ideal. Departures fall into four natural categories:

1. Incomplete lists of schools
2. Nonparticipation by schools
3. Incomplete lists of students
4. Nonparticipation by students

A comprehensive research plan must focus on each of these, assessing the extent of departure from the ideal in each case, as well as the impact of those departures on population achievement estimates. The impact of the departures is particularly important when it affects comparisons. For example, if between 1998 and 2002 there is an increase in the percentages of students with disabilities who do not participate in NAEP, then unless there is appropriate correction for that increase, bogus reports of achievement gains between 1998 and 2002 will result.

Incomplete Lists of Schools

For sampling, NAEP has made use of lists of schools created by the National Center for Education Statistics (NCES) and by Quality Education Data, Inc. At NCES, the Common Core of Data (CCD) identifies K–12 public schools in the nation, and the Nonpublic School Universe Survey identifies K–12 nonpublic schools.

CCD contains many entities other than regular schools, and NAEP limits the public school assessments to regular schools. Although the information in CCD, as reported by the State Education Agencies, is quite comprehensive, there are nevertheless imperfections in the data. Errors in the specification of grade ranges and delays in the addition of new schools to the CCD database are two ways in which the school lists for NAEP might be inaccurate. Inaccuracies in the nonpublic school lists are more likely than in CCD because there is no single national framework for identifying nonpublic schools.

Even if there were no errors in the sampling frames, there are still departures from the representation of schools containing *all* students. Schools for blind and deaf students, schools in correctional facilities, and home schools are examples of schools not included in the NAEP school sampling frame.

These imperfections have been tolerated in NAEP because they do not detract from the intent to focus on the mainstream of education *and* because the percentages of students enrolled in these excluded schools is so small that they cannot have a noticeable effect on state or national averages.[4] If there should be a growth (or decline) of any type of school not included in the NAEP sampling frame, or a state in which a much larger than average percentage of students are in such schools, then the imperfections might become problematic. Monitoring general trends in non-regular school enrollments across states is a relatively inexpensive activity that can guard against unfortunate surprises.

---

[4] In 1999–2000, according to preliminary CCD figures, 98.2 percent of students in public schools were in "regular" public schools.

Nonpublic schools are much more likely than public schools to open and close, and the task of ensuring a complete nonpublic school sampling frame is significant. This is a direct problem only for national NAEP, because state NAEP specifically focuses on students in public schools. However, if a state experiences an unusually high movement of students to or from private schools, that can distort public school achievement trends.

The priority of research on this validity issue depends on the size of the effect on NAEP results. While the percentages of students in non-frame schools remain small, validity research on this topic may reasonably be limited to a) monitoring of student flows in participating states to identify points of significant growth that would warrant consideration of expanding the sampling frame and b) sensitivity analyses to indicate the size of biases that might be present if students in excluded schools were to score a standard deviation above or below students in schools in the sampling frame.

In 2001, there are two major growth sectors of public schools: charter schools and outplacement facilities. Charter school growth is sufficiently widespread to raise concern that this category of schools not be omitted from the NAEP sampling frame. While there is no *a priori* expectation that students in charter schools will score higher or lower than students in regular public schools, the nature of charter schools in different states may differ, and charter schools in some states may be more likely than in others to be included in the NAEP sampling frame. Therefore, student movement from other schools to charter schools can affect the measurement of achievement trends. A comparison of the NAEP sampling frame with other lists of charter schools, such as that developed in U.S. Department of Education studies of charter schools is warranted.

Students with special needs who are in outplacement facilities are increasingly being included in testing. As their numbers increase, there is an increasing threat to the accuracy of NAEP statistics that ignore students in these facilities. Since these students are normally educated with funds that are allocated to the school that "outplaces" the students, it might be possible to ensure that these students are included in the student lists for selected regular public schools. However, the student assessment data will be more useful if accompanied by data on the resources and context of the outplacement facility, as a school. Therefore, it may be better to include these facilities in school lists. Validity research is needed that focuses on finding the most efficient method for including these (very small) schools. Of course, many of the students in outplacement facilities have disabilities that limit their participation in tests, a topic which is discussed below in the section on excluded students.

Non-Participation by Schools

Participation in national NAEP, and in many states in state NAEP, is voluntary and at the discretion of local schools and districts. Because there is concern that the schools that refuse to participate might have students who would perform differently from students in participating schools, NAEP takes two important steps to maximize the representativeness of the set of participating schools. First, a set of substitute schools is carefully selected to replace schools that cannot be persuaded to participate; and second, states which fail to meet a criterion of 70 percent original participation and 85 percent after-substitution participation are excluded from NAEP reporting.

The important question remains as to whether the substitute schools are sufficiently similar to the refusing schools they replace that do not introduce bias in the results. A study of this bias was undertaken in conjunction with the 1994 state NAEP. That study compared the states' own assessments of students in the two sets of schools as indicators of school performance and found no significant differences in the mean performance of NAEP's refusing and substitute schools. Such a study requires the acquisition of school-level state assessment data from individual states; with the availability of such data, the analyses to test the hypothesis of no difference are simple and straightforward.[5]

Since AIR has collected these data for other purposes, the studies should be carried out as a part of the routine documentation of state NAEP. For national NAEP, which includes nonpublic schools, the acquisition of comparable achievement data from refusing and substitute schools is not as simple. Nevertheless, since national NAEP produces policy-relevant comparisons between public and nonpublic sectors, a periodic study of replacement in national NAEP is warranted.

Incomplete Lists of Students

Lists of enrolled students are provided to NAEP by participating schools in October and November, and NAEP is administered in February and March. To control for the possibility that students who change schools between November and February perform at lower levels, on average, than students remaining in the same schools, NAEP asks schools to add to their testing sessions a supplementary sample drawn from the students who were not included in the original sampling list but who are enrolled at the time of the NAEP administration. An assumption is made that the students added in the supplementary samples are representative, in aggregate, of the originally selected students who failed to participate because they left the schools in which they were listed before the NAEP administration.

There are other ways in which the student sampling frames in schools can be incomplete. These include: (1) students not counted in the regular enrollment in November; (2) students (for example, in year-round schools) who change grades between November and February; and (3) students in ungraded classrooms. The sizes of these different categories of students are unknown, and since the performance of these students may well be different from the performance of listed students, a thorough review of the sampling frames in a sample of schools may be needed.[6] A first stage would involve estimating the numbers of students omitted from the sampling frames for various reasons. A second stage, finding out how the performance of the different categories of students varies, will only be necessary if the numbers are large or growing.

---

[5] In fact, such a study could easily be extended to the estimation of the effects of a) school refusals that are not replaced and b) the actual sampling error due to the sampling of 100 schools to represent all schools in the state.

[6] Anecdotal evidence indicates that in year-round schools, students enrolled in February tend to be no different from other students (parents generally want uniformity over the year). However, for some students in these schools, February may be the "September" of their school year.

Non-Participation by Students

NAEP creates an Administration Schedule for each testing session listing the students expected to participate in that session. Differences between these lists and the records of student performance in the resulting NAEP database can arise either because students are excluded from participation by the school or because students fail to appear for the testing session. Both of these categories may distort estimates of achievement, and NAEP has taken steps to minimize each. Nevertheless, research is needed to ensure that these processes do not bias NAEP results.

Students Who Fail to Appear

First, the typical NAEP testing session involves 25 to 30 students, and NAEP test administration instructions indicate that if as many as five students fail to appear, then a make-up session is to be scheduled so that most or all of the absent students can be tested. Thus, absences are rarely more than 10 percent of the selected students in any school. In combining the assessment results to estimate achievement for a population, the scores of other students, similar to the absent students, are imputed to the absent students.[7]

Other students who fail to appear are the students who left the school between November and February. These students are excluded from the population estimation process, but in order to include mobile students in the estimates, the supplementary sampled students (some of the students who arrive at the school between November and February) are included in their place. This is an adequate adjustment in the aggregate if the flow of students between schools is uniform over time. However, it would fail to include students who are "between schools" from November to February. These include foreign students who go home for extended Christmas vacations.

As long as students only move between schools that have a positive probability of being included in the NAEP sample, the supplementary sample of incoming students is an adequate replacement for students who withdraw from schools. However, any tendency for poorer performing students to move out of the system (e.g., by dropping out, or in state NAEP, moving to private schools or schools in other states) and better performing students to move into the system (e.g., from abroad) in the middle of the school year, or vice versa, would introduce bias in the estimates.[8] A careful comparison of withdrawn students and supplementary students in one operational state NAEP could dispel a great deal of uncertainty about this potential source of bias.

---

[7] At present, the method of imputation is by reassigning the "weight" of absent students in the aggregation as an addition to the weight given to other students with scores. A more sophisticated form of imputation, making use of multiple imputation methodologies and using extended information about the performance level of absent students, could improve the accuracy of population estimates.

[8] Whether the loss of within-year dropouts is a bias is a conceptual issue. If the NAEP population is "all fourth graders enrolled in February," then dropouts prior to February are irrelevant. However, there is a sense that "fourth graders in the state" refers to a population that includes students who enrolled in fourth grade in September. Schools are responsible for their students for the full school year, unless the students transfer to other bona fide schools. The specification of the NAEP reference population should be based on the utility of the definition, with technical solutions to the problems of assessing the defined population, not the other way around.

Students Who Are Excluded

On the student lists are indicators for each student as to whether he or she has a disability (labeled in this discussion as an "SD student") and as to whether he or she has limited English proficiency (labeled here as an "LEP student"). If a selected student is either SD or LEP, the school may determine that the student cannot meaningfully participate in NAEP and, therefore, exclude him or her from the testing session. Over the years, NAEP has taken no step to represent those excluded students in population estimates but has tried various criteria and instructions to ensure that the same kinds of students would be excluded in each state, district, and school.

Until the mid-1990s, this was not problematic because only a small constant percentage of students was excluded. However, since 1995 there have been significant increases in the percentages of students excluded from NAEP, and these increases have been shown to distort NAEP's estimates of population gains. The major reason for the increases is, ironically, the federal legislation that *all* students should be included in testing programs. As a result of the inclusion effort, state assessment programs have developed policies of "accommodating" testing conditions for students with disabilities or limited English proficiency. For example, children with learning disabilities may be given extra time or help with reading the test items. It cannot be doubted that these accommodations affect the scores of children, and some state testing programs have implicitly acknowledged that fact by publishing scores for both the general population and the full population. However, because there is no firm research foundation for estimating the size of effects of accommodations on test scores, NAEP has limited the use of accommodations. Since local school policies dictate that SD and LEP students should be accommodated in testing, the result is that they are excluded from NAEP.

Because NAEP collects extra information on SD and LEP students and includes about half of these students in testing, grounds exist for imputing the performance distribution of excluded SD and LEP students to use in generating full population estimates. The use of such an imputation to estimate state NAEP mathematics gains from 1996 to 2000 is shown in the appendix of this paper. The assumption underlying this imputation is that included and excluded SD and LEP students would perform the same in each state on average, except as indicated by differences in background information and SD/LEP questionnaire responses. The relations of these measures to performance is estimated based on their correlations with performance of included SD and LEP students.[9]

As an alternative, NAEP is moving to the strategy of including test scores based on accommodated administrations of NAEP to minimize the percentage of students who are excluded from NAEP population estimates. However, at present this strategy is flawed because there are no firm estimates of the effects of accommodations on performance scores. Moreover, the determination of what accommodations a student is provided is not yet well-controlled, so many students who would have participated in NAEP without accommodations will receive accommodations in the future. Thus, publishing estimates based on different mixtures of accommodations for different groups (e.g., in different states or in the same state at different times) confounds any attempts to make inferences about comparisons.

---

[9] More specifically, the relations are estimated from pooled within-state linear regressions predicting NAEP plausible values from information on the SD/LEP questionnaire, on the Administration Schedule (race and gender), and school-level factors (such as poverty level). These measures account for approximately 30 percent of the within-state variance in NAEP plausible values of included SD and LEP students.

Although improved SD/LEP questionnaires can increase the precision of imputation of scores of excluded students, the "problem" of accommodations must ultimately be resolved by research to estimate the effects of accommodations on test scores. The most urgent validity research need for NAEP is to fill the information need for estimates of the effects of accommodations on performance. Once those parameters are known, full population estimation will be less sensitive to variations in accommodation policies.

The design of the needed accommodations research is to gather distributions of scores of students on the "same" test randomly assigned to accommodated and non-accommodated conditions. The scope of the study must include all the kinds of students who might possibly be accommodated in NAEP, and it must include sufficient numbers of students to support the estimation of effects precisely enough to enable NAEP to validly aggregate scores from accommodated and non-accommodated administrations.

Planning for research on accommodations is complicated by the accompanying focus on the definition of the domain being measured. Arguments are made that the nature of what NAEP is testing should be changed by implementing accommodated administrations for all students. For example, in one state items on the state's "reading" assessment are read to students who need it, on the grounds that the skill to be assessed should be not reading printed text but comprehending linguistic communication. These arguments cannot be resolved scientifically because they involve judgments as to what skills are "important." Because these debates can be expected to continue far into the future, decisions must be made to conduct accommodations research in the context of reporting achievement according to a standard definition of reading, mathematics, science, and writing performance. The use of the current (unaccommodated) NAEP administration for that standard has the advantage that the results of the research will aid the reporting of NAEP achievement trends from the past thirty years to the future.

Establishing Research Priorities

Three types of validity research are needed: conceptual, analytical, and empirical. For the purposes of establishing the appropriate definition of the target population for NAEP estimation, for example, a conceptual consensus is needed. A "population definition" commission representing NAEP constituencies must compare the value of NAEP under different target population definitions and establish a single definition that will then guide sampling, instrument development, and statistical estimation procedures. While such a commission must be informed by information on the cost and feasibility of estimating performance of alternative populations (e.g., including dropouts and home schooled children may be infeasible), the utility of the data should be the paramount consideration.

A great deal of analytical work can be done by a) using NAEP and other assessment data already available, and b) using simulated data to explore hypothetical situations. At any time, one can embark on analytical studies to estimate the impact of a variety of departures from the ideal sample-to-population relation on comparisons between groups, between states, and over time. For example, the analyses presented in the appendix of this paper are based entirely on state NAEP data.

Empirical work is, with exceptions, more expensive than analytical or conceptual research. Searching sampled areas for entities providing instruction to students not enrolled in regular schools involves substantial staff time to examine directories, conduct telephone interviews, and make site visits. Similarly, comparing NAEP student lists with the characteristics of actual students attending selected schools requires firsthand observation of students in schools. And of course, research on accommodation, which might be considered "instrumentation research" rather than "sampling research," involves a great deal of new data collection.

A reasonable starting point for a program of NAEP sampling validity research would be a matrix crossing the four types of threats to validity (school and student lists and school and student non-participation) with the three research types (conceptual research, analytical research, and empirical research). The next step is to establish priorities for the various components of the research program.

The selection of research for funding should be based on comparisons of cost, benefit, and timing. The first step is to estimate the potential benefit of the research in terms of elimination of threats to the validity of NAEP publications. (That estimation is, itself, an analytical and conceptual research project that should be undertaken immediately.) The threat of a particular type of departure (such as student absences) from ideal full population estimation is a function of both the frequency with which "errors" (e.g., absences) occur and the impact of those errors on important comparisons.

The most important fact about research costs is that the cost of research to address a topic can be adjusted within broad ranges to fit funding constraints. With that perspective, NAEP can make priority decisions about research funding based on the precision with which answers to questions need to be known. For example, if nothing is known about the prevalence of home schooling, a one- or two-week effort might suffice to acquire "ballpark" information about state testing policies and enrollment estimates for home schooled students. On the basis of such a study, the estimates needed for deciding how much to allocate to a larger study can be generated. The smaller the numbers of such students, the less that needs to be known about how their academic achievement differs from the achievement of students in regular schools.

An effective method for minimizing the costs of validity research studies is by packaging them. For example, many validity studies in the past made use of data already collected as a part of either an operational NAEP administration or a NAEP field test. Another example is that conceptual projects to clarify definitions and priorities can both benefit from analytical work and serve to guide further research. Likewise, studies that involve similar operations can be integrated, such as by conducting a single set of case study visits to assess the completeness of both school lists and student lists.

Because any one of the validity threats discussed in this paper has the potential for damaging NAEP, this perspective on costs leads to the recommendation to proceed on all fronts—to invest at least some effort in each of the areas, and to use initial results to focus on the areas in which the need for information to remove the threat is greatest.

Finally, there is the issue of the timing of research. In a steady state, there should not be an urgent need for ongoing validity research, that is, for replication of studies previously carried out. However, when the context of NAEP changes, urgent needs suddenly appear, and they tend to disrupt ongoing processes. For example, the increasing use of accommodations in state assessments has created a largely unanticipated problem for NAEP. A significant effort should be allocated, on an ongoing basis, for anticipating the effects of the changing environment on threats to the validity of NAEP. In 2001, there are a variety of issues surrounding the possibility of NAEP's taking on a high stakes role in federal funding of education, and NAEP is focusing on these. There are also other changes in the current environment that can threaten the validity of NAEP, including redefinitions of public schools, such as charter schools, schools within schools, and year-round schools. These trends should be monitored to determine how they might affect the validity of NAEP.

Chapter 5. Issues and Recommendations on NAEP Data Analysis

Subcommittee: *R. Darrell Bock*
Albert Beaton
Gerald DeMauro

Present procedures of NAEP data analysis for results reporting are essentially the same as those formulated by Mislevy and others in 1984. Although they have served the assessment well in the intervening years,[1] the excessive amount of time required for completing the analysis has long been a contentious issue.[2] In this paper, we address the question of how the elapsed time between the completion of data collection and the initial report of results could be shortened by simplifying the methods of data analysis. I am indebted to Eugene Johnson for identifying the following sources of delay attributable to the statistical analysis of the data:

- Present procedures do not distinguish between the preparation of results required for the primary report of achievement levels in the general population and main subpopulations from those that serve the secondary purpose of relating the results to various sociological and psychological variables derived from the student, teacher, and school questionnaires. As a consequence, all data cleaning including that of the questionnaires must be completed before the analysis can begin. Because of limits on the number of persons that can be assigned to this task, the excessive time required for data cleaning appreciably delays the analysis step. In addition, if any error is found in the coding of questionnaire responses after analysis, a complete reanalysis is necessary.
- Present procedures are oriented toward the production of a single master result file that will serve the purposes of both primary achievement results reporting and secondary analysis by both intra- and extra-mural groups working with (or potentially working with) the assessment results. For each examinee participating in the national or state samples, this file contains multiple imputations of so-called *plausible values* for each subject matter variable in a particular assessment. The so-called *conditioning* step required to produce this file uses variation in all background data—several hundred distinct pieces of

---

[1] Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.
Mislevy, R. J., Johnson, E.G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131–154.
Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161.

[2] Bock, R.D. & McCabe, P. (1990). *Toward a timely state NAEP*. Report to the National Assessment Governing Board.

information—to predict the mean of a prior distribution of proficiency for each particular examinee. The examinee's responses to the exercises on a given test form then provide the remaining information that determines his or her posterior distribution of proficiency. Once the six primary reporting variables (race/ethnicity, type of location, parents' education level, region of the country, modal age for grade, type of school) have been included in these predictions, however, the marginal improvement in predictive accuracy attributable to the remaining background variables is relatively small. Delaying results for the primary reporting variables while the remaining background variables are prepared for analysis is therefore difficult to justify.

- The fact that achievement measures for the subject-matter areas of science, mathematics, and reading are reported both as subarea scores and overall scores greatly complicates the conditioning process. Further complexity is introduced into the analysis by the definition of the overall area score as an arbitrarily weighted composite of subarea scores.
- Finally, a source of delay not associated with measurement issues is the need to await the calculation of post-stratification weights before the final analysis can begin. Past experience suggests, however, that the effect of changes in the weights from their pre-stratification values are typically too small to influence any policy-relevant inferences that might be based on a provisional report of primary results prior to reweighting.

Expediting NAEP Primary Results Reporting

In the nearly 20 years since the present analytical procedures of NAEP were developed, innovations in item response theory (IRT) and statistical methods have occurred that can potentially simplify and speed the analysis of NAEP data without compromising the integrity of the reporting scales or trend statistics. Briefly described, the contributions of these methods to solutions of the above problems are as follows:

- Multiple-group IRT permits parameter estimation for a defined set of items to be carried out simultaneously in groups of examinees sampled from more than one sub-population.[3] By the method of maximum marginal likelihood (MML), the proficiency distribution in each group is estimated jointly with the item parameters. From the corresponding latent distributions, the means and standard deviations of proficiency in each group, and the percent of examinees above any specified achievement-level criterion (PAC), can be estimated directly. Estimation of scores or plausible values for individual examinees is not required.
- A similar MML procedure that estimates regression relationships among examinee background characteristics as well as means, standard deviations, and PACs has been formulated by Cohen and Jiang.[4] It is suitable as an analytic

---

[3] Bock, R.D. & Zimowski, M.F. (1997). Multiple group IRT. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer.

Bock, R.D. (1989). Addendum—measurement of human variation: a two-stage model. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 319–342). New York: Academic Press.

[4] Cohen, J. & Jiang, T. (2001). *Direct estimation of latent distributions of large-scale assessments with application to the National Assessment of Educational Progress (NAEP).* Washington, DC: American Institutes for Research.

method for preparing NAEP reports and for other research groups to use in analyzing NAEP data in greater detail.

- Item bifactor analysis provides efficient estimation of a general factor of proficiency in an item set that also represents two or more special proficiencies.[5] The general factor standard error of estimation correctly accounts for failure of conditional independence within the item set. The latent distribution of the general factor is available for estimating group means, standard deviations, and PACs. The procedure allows straightforward analysis of any NAEP subject area that is reported both as subarea proficiency and overall proficiency.

Proposed applications of these procedures to item parameter estimation and primary and secondary results reporting are described in the following subsections.

Item Parameter Estimation and Forms Equating

The main point at which the present proposals depart from the existing NAEP procedures is in restricting the conditioning of item parameter estimation to five or six of the NAEP primary reporting variables. According to Appendix B of the 1996 Technical Manual,[6] the categories of the primary reporting variables are as follows:

1. Race/Ethnicity
   - Black
   - Hispanic
   - Other
2. Type of Location
   - Central City
   - Urban Fringe/Large Town
   - Rural/Small Town
3. Student's Report of Parents' Education Level
   - Not finished high school
   - Graduated high school
   - Some education after high school
   - Graduated from college
4. Region of the Country
   - Northeast
   - Southeast
   - Central
   - West

---

[5] Gibbons, R.D. & Hedeker, D.R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

[6] Allen, N.L., Carlson, J.E., & Zelenak, C.A. (1999). *The NAEP 1996 technical report.* Washington, DC: National Center for Education Statistics.

5. Modal Age for Grade
    - Younger than modal age
    - At modal age
    - Older than modal age

6. Type of School[7]
    - Public (including BIA and DODEA schools)
    - Private (including Parochial Schools)

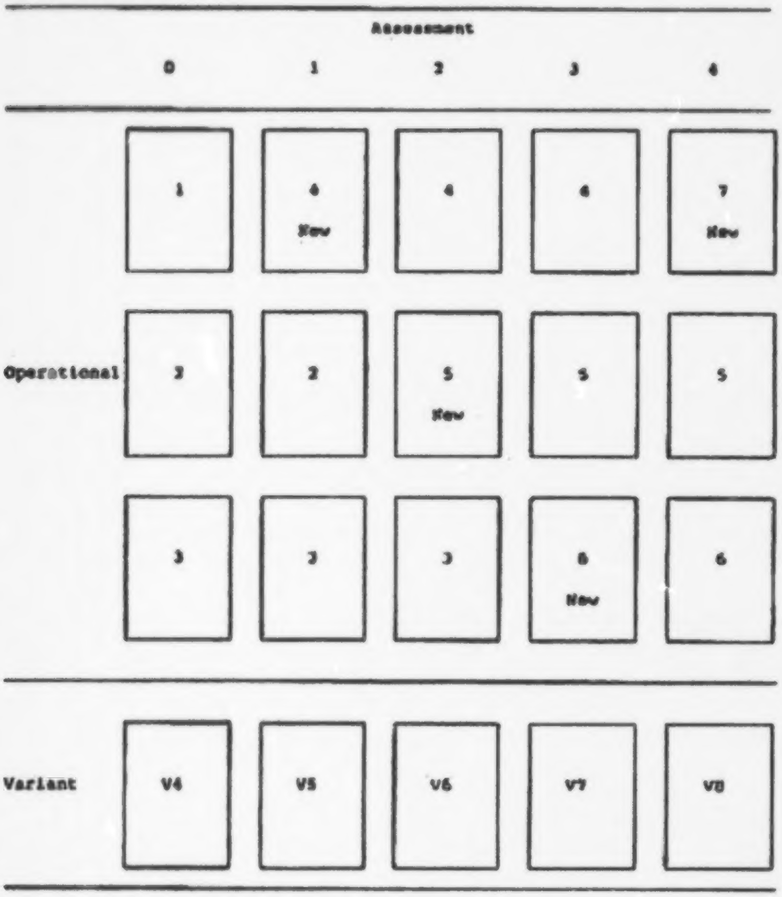
These variables acting jointly are among the strongest predictors of achievement level. The proposal is to condition the item parameter estimation and equating only on the effect of the subgroups generated by the cross-classification of examinees with respect to these variables. This results in 864 subgroups for Main NAEP and 432 for State NAEP. If item parameter estimation is then carried out by multiple-group IRT for all non-empty subgroups, the prior distribution of proficiency required in MML estimation is individualized to whichever subgroup the examinee belongs. With conditioning limited in this way, the analysis of primary results reports can be carried out as soon as the data are cleaned and certified for these variables alone. Cleaning of the remaining data, especially the remaining questionnaire responses, can then continue in preparation for a later secondary analysis report.

The procedure for combined item analysis and forms equating is predicated on a continuing program of forms development in which one-third of items in the previous form are replaced in each new form. As an example, Figure 5.1 shows schematically how the item composition of three forms in a particular subject area is renewed through five waves of assessments. Note that each assessment includes a set of so-called "variant" items, which are being readied for inclusion in the following assessment, but do not enter into the item calibration of the current assessment. Typically, the variant items number somewhat more than one-third of those in the assessment proper to allow for culling of items that have unsatisfactory information characteristics.

---

[7] This variable is not included in State NAEP.

**Figure 5.1–
Item-set Composition of the First Five Waves of a Continuing Assessment Replacing One-third of Items Each Wave**



In an extensive simulation study, Hedges and Vevea recently examined various methods of forms equating that apply to assessment schemes such as that in Figure 5.1.[8] They found the best performing of these methods to be a multiple-group approach. In its simplest form, this method of equating makes use of item response data from the two most recent assessment years containing common items. The data for each year enters the analysis as a distinct group with its own latent distribution for the subject area being measured. (Although there are typically multiple test forms in each assessment year, there is only one *test*, defined by the union of all items on all non-variant forms.) During item parameter estimation, the means and standard deviations of the latent distributions corresponding to the two assessment years are concurrently estimated. As usual, overall location and scale of the latent distributions are indeterminant. Determinacy is resolved by setting the mean and standard deviation of the previous assessment year to the values that were estimated and reported at that time. This determines location and scale in the current year on the scale of the previous year and thus equates the forms. Change in achievement levels may then be expressed as between-year differences for the various demographic groups represented in the two assessments. This procedure has the merit of using all information available in the data from the two years in estimating the item parameters and the group means and standard deviations.

---

[8] Hedges, L.V. & Veva, J.L. (1997). *A study of equating in NAEP*. Palo Alto, CA: American Institutes for Research.

A second method of forms equating investigated by Hedges and Vevea makes use of data only from the current year.[9] Parameters of all items carried over from the previous assessment are fixed at the values previously assigned to them. In the analysis, item parameters are estimated only for those items appearing in the assessment for the first time. With location and scale determined by the fixed parameters, all estimates of demographic group statistics are automatically on the already existing scale and differences between the assessment years are interpretable. In terms of accuracy, this method is sub-optimal relative to the multiple group method, but the simulation results showed the loss to be rather small. Either method is a worthy candidate for forms equating in an operational assessment.

Estimation of Sub-population Latent Distributions Given the Item-parameter Estimates of the Current Assessment

For purposes of characterizing achievement performance of any demographic group within the national sample, the latent distribution of examinee proficiency may be represented as weights on, say, 80 equally spaced points between –4.0 and +4.0 on the proficiency dimension. Given the values of the item parameters and the item response pattern of a particular examinee, the posterior probability density can be evaluated at each of these points provided one makes some assumption about the prior distribution of proficiency in the group to which the examinee belongs. The conventional assumption is that the examinee is drawn from a normal distribution with the mean and standard deviation estimated from the data available from all examinees in the group. The estimation procedure is carried out iteratively starting from some provisional value of the mean and standard deviation. In each iteration, posterior probabilities at the 80 points are aggregated case-by-case for all examinees assigned to that group. When the aggregation is complete, the weights are normed to unity by dividing by their total, and the mean and standard deviations of the distribution represented by the weights are calculated. These values provide the parameters of the provisional prior distribution for the next iteration. The iterations continue until the estimates become stable.

With the latent distributions of multiple demographic groups estimated in this way, their aggregate distribution can be calculated by summing the weights at each point multiplied by the number of cases in the group. Using the usual formula for grouped data, means variances and higher moments of the distribution can then be computed. The achievement-level PACs can be obtained simply by summing the weights up to the point below criterion and interpolating if the criterion point falls between successive points on the continuum.

In the case of the five or six primary reporting variables, the latent distributions of all sub-categories in the classification of cases are by-products of item parameter estimation and do not have to be recomputed. Higher-order margins for any of the variables can be aggregated by summing over the distributions of the ignored variables. Even in the full national sample, however, some of the subgroups will undoubtedly be empty, or nearly empty. If the subgroup is actually empty, its latent distribution does not appear in the marginalization. If there is even so much as one case, however, there will be a profile of posteriors over the 80 points, but it will have very little influence on any higher-order margin.

---

[9] This is essentially the method currently in use by ETS (J. Donoghue, personal communication).

This method of estimating group characteristics is very similar to the present procedure based on aggregation of plausible values. Apart from the reduced conditioning, the only material difference is that this method is based directly on latent distributions and does not assume that the examinee's posterior distribution of proficiency is normal. The direct procedure will therefore be somewhat more accurate, especially when the number of items in the response pattern is small and posteriors are unlikely to be symmetric, let alone normal.

For reporting variables other than the primary ones, the data must be passed through one more time to obtain the empirical priors for the demographic categories and accumulate the latent distributions. The item parameters are fixed at the values obtaining the analysis for the five or six primary reporting variables. Once the latent distributions for the categories of these variables are computed in the multiple group analysis, they may be aggregated to higher margins in the same manner as the categories of the primary variables.

This direct approach to generating reporting statistics for the assessment can be extended to the estimation of regression relationships among the primary and secondary reporting variables. A computer program for that purpose has been prepared by Cohen and Jiang.[10] If that program, along with the estimated item parameters and the original item response data of the assessment, are made available to secondary users, these users can perform directly all of the analyses they might have performed with the plausible values presently supplied in the NAEP secondary users file. This includes estimation of PACs by the same method as in multiple-group analysis. In simulated data and in data from the 1992 and 1996 assessments, Cohen and Jiang found good agreement between direct estimates of primary reporting statistics and those obtained from plausible values.

The item parameter estimation and equating phase of the proposed analytical procedures can be carried out with the BILOG-MG program of Zimowski, Muraki, Mislevy, and Bock.[11] This program handles multiple tests and multiple test forms, as well as the multiple groups, and is suitable for complex, large-scale testing programs such as NAEP. The present version of the program is limited to binary scored items, however, and would have to be extended in order to include the polytomous scored items of the NAEP assessment instruments.[12]

---

[10] Cohen, J. & Jiang, T. (2001)

[11] Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1995). BILOG-MG: multiple-group item analysis and test scoring. Chicago: Scientific Software International.

[12] This method can be implemented using NAEP's operational software (J. Donoghue, personal communication)

Cross-checks on the Forms Equating

Because any measure of change from one assessment to another is absolutely dependent on the validity of the equating, it is important to have cross-checks on equating validity. For this purpose, a provision in the assessment operations should be made to have a certain number of current and previous year's forms assigned randomly to examinees at a number of assessment sites. With sufficient data from such sites, one may then perform random equivalent groups equating that requires less stringent assumptions than the common item equating described above. Cross-checks between the two types of equating should be a routine part of every assessment.

Where open-ended exercises are involved, another important cross-check is on the consistency of rating procedures from one assessment to another. To the extent that rating procedures have a subjective element and require training of rating teams, it is always possible for the definitions and standards to change somewhat as new raters are recruited and trained. For this reason it is essential that the new teams grade a certain number of tests from the previous assessment blindly. In this way, any systematic change and severity or leniency of ratings can be detected and corrected for. See Bock for a discussion of this type of control procedure as applied in the Direct Writing Assessments of the California Assessment Program.[13] For a discussion of the standard errors of IRT scale scores in the presence of multiple ratings see Bock, Brennan, and Muraki.[14]

Other Issues: Subject-matter Subarea Scores

A remaining troublesome problem is the procedure for reporting of subject-matter proficiency measures expressed as weighted composites of more detailed subarea measures. At present, the weights for the subareas are arbitrarily assigned by the subject-matter content committee. This problem arises, for example, in the science area, where the detailed subareas are Physical Science, Biological Science, and Earth Science. If the test form presented to a given student contains only one subarea, separate posterior latent distributions for each subarea could simply be summed (weighted sum) over subjects to obtain the latent distribution of the composite measure. If two or more subareas are represented in each form, however, and the subareas represent different but correlated latent dimensions, the sum of the posterior distributions for a given student would not equal the posterior for the subarea items treated as one-dimensional for reporting purposes.

The problem of more than one subarea per form would not arise, of course, if the items were treated as a single test for purposes of reporting overall performance in the subject area. The proportion of distinct items in each subarea of the assessment instrument could then be set to reflect roughly the weights assigned to the subareas. The subject-matter score would add another measure for which group latent distributions would be calculated along with those of the separate subareas.

---

[13] Bock, R.D. (1995). Open-ended exercises in large-scale educational assessment. In L.B. Resnick and J.G. Wirt (Eds.), Linking school and work: Roles for standards and assessment. (pp. 305–338) San Francisco: Jossey-Bass.

[14] Bock, R. D., Brennan, R. L. and Muraki, E. (in press). The information in multiple ratings. *Applied Psychological Measurement.*

A more exacting approach could be to make use of a quasi-likelihood in the IRT analysis in which weights are assigned to the log likelihood of each item, so that the average information contributed by each subarea matches the weight assigned by the content committee. This would require a preliminary analysis to determine the unadjusted average information of the subarea items. However, if the subarea proficiencies are highly correlated, which seems likely, the effect of the weights on the overall performance measure will be small whatever method is used.

If the weights have a large effect, it is prima facie evidence that the subareas measure different latent dimensions and that the conditional independence required in unidimensional IRT analysis is violated. In that case, the best approach would be to estimate the overall measure using the Gibbons and Hedeker item bi-factor model cited earlier. The two-dimensional MML IRT analysis under this model yields best estimates the general factor common to the subareas and gives the correct standard error of estimate of the student's proficiency on the general factor. Comparison of this standard error with the smaller (erroneous) error of the unidimensional solution indicates the extent of the failure of conditional independence.

An intermediate result in MML estimation for the bifactor model is the marginal posterior latent distribution for the general factor. This distribution can be used to estimate group means and standard deviations, as well as PACs, in the same manner as the posterior latent distribution of unidimensional IRT models.

Other Issues: The School as the Unit of Analysis in Studying Relationships Between Demographic Variables and Educational Outcomes

In place of the plausible values conditioned both on reporting variables and background variables, NAEP could provide secondary investigators with much the same information in a public use file containing directly estimated school-level results. That possibility was not explored in the formulation of the 1984 analysis plan because, at that time, a fundamental assumption was that the student should be the basic unit of data analysis. Schools from which students were sampled were to play no role in the analysis and were not to be identified in any way in the reports or in the public use tapes. Existing legislation protecting the privacy of information held by government agencies was assumed to apply to schools as well as to individual students. It was also thought that school officials at the state and local levels would not agree to the participation of their schools without this anonymity. Background data descriptive of each school were collected, of course, but in the public use tapes they were attached to the student records without distinguishing between schools.

With the growth of the accountability movement, however, attitudes toward reporting of information about schools have changed dramatically. Most public school systems submit annual test results to the media for publication, post them on the Internet, and invite comments on the implications of the findings. In some states, evaluation of schools based on test results plays a major role in awarding funds for program improvement and in initiating or monitoring the reform of poorly performing school systems. In these states the availability of confidential NAEP results identifying the schools in the sample would provide an independent "spot check " on the dependability of the state test results for school-level decisions. In addition, secondary analysis of school-level reports of

NAEP achievement and background questionnaire data, even without identification of individual schools, has the potential to provide guidelines for these high-stakes uses of test results.

There is little likelihood in the present climate of opinion that anyone would object to distinguishing schools in a public use file that contains only school-level data. Records in this file would consist of direct group-level estimates of student proficiency and their standard errors, responses to the school questionnaire, and school-level aggregates of responses to the student questionnaire. No codes or data that would identify individual schools would be included in the secondary users' file. Achievement Results in this file would be disaggregated below the school level with respect to only one variable—namely, gender of student. That is, means and standard deviations of test scores for each school would be reported separately for each gender. Aggregate information on other categorical attributes of individual students would be conveyed by their frequencies among the students sampled within the school.

Having NAEP data available in this form would be of interest to secondary users who wish to study the relationships among community and neighborhood characteristics represented in the background questionnaires. They would find the mean proficiency measures attractive because these measures have sufficiently small standard errors to be treated directly as observations for weighted least-squares analysis (with the reciprocal squared errors as weights). The aggregated questionnaire results could be expressed as frequencies and accumulated to various margins for use in log-linear analysis. Or they could be expressed as logits to serve as variables in regression analysis or structural equation modeling. Inference in these analyses would be directly to the population and subpopulations of schools. For studies of educational policy issues, these inferences are often more relevant than inferences about populations of students. Also attractive for secondary users would be the compactness of the data. The number of records in the file is the number of schools—between one and two orders of magnitude smaller than the number of students in a NAEP sample. NAEP's own analyses and reports of background variable effects could, for the most part, also be based on school-level summaries of student performance.

The school-level proficiency estimates could be obtained by extending the multiple-group IRT model to include the schools as a random way of classification. In other words, there would be three levels of random effects: responses within examinees, examinees within schools, and schools within higher-level fixed groups. Each school would have a prior distribution within a higher-level fixed group (for example, "Type of Location" or "State"), and within each higher-level group there would be a posterior latent distribution of schools with the mean and standard deviation descriptive of the fixed group. In the scoring phase of the analysis, Bayes or maximum likelihood estimates of school-level proficiency would be computed along with their corresponding posterior standard deviations or standard errors. These estimates would be treated as observations in school-level secondary analyses. Note that in small schools where the number of students required in the NAEP sampling design cannot be met, the stability of the results would still be assured by the effect of the empirical prior distribution of schools within fixed groups in strengthening estimation with sparse data.

At a minimum, school level summaries would be merely another way of reporting existing NAEP data. Present reports by demographic groups at the state, regional, and national level would not be affected. If interest in the school level reports should become more salient, however, the NAEP design might move toward greater numbers of students per school and collection of more detailed school and community information.

Other Issues: Including Item Sampling in NAEP Jackknifed and Standard Errors

Because of the complexity of the NAEP data analysis and sample structure, the standard errors of the reporting statistics are estimated empirically by Tukey's jackknife procedure. These standard errors represent the uncertainty of inference from the sample of students in the assessment to the population of students in the reporting categories. There is, however, another domain of sampling that cuts across the sampling of students--namely, the sampling of replacement items in successive assessments of the same content area. Since there is always uncertainty involved in generalizing from the items in the assessment instrument to the domain of items defined by the subject-matter content, the replacement of a substantial proportion of items each time a given subject area is assessed is an important source of error variation that is not reflected in present NAEP standard errors.

Recently, Cohen, Johnson, and Angeles have applied the jackknife procedure jointly with respect to sampling of students and sampling of items.[15] This work makes the estimation of standard errors in NAEP consistent with the generalizability concept of modern test theory. For many of the reporting variables, the authors found the component of error variation attributable to item sampling to be as large or larger than the component attributable to sampling of students. This source of uncertainty in NAEP achievement scores needs further investigation.

---

[15] Cohen, J., Johnson, E. & Angeles (undated). *Variance estimation when sampling in two dimensions via the jackknife with application to the National Assessment of Educational Progress.* Washington, DC: American Institutes for Research.

Chapter 6. Validity and Utility Issues in NAEP Reporting and Data Releases

Subcommittee: Frances B. Stancavage
Ina V.S. Mullis

Ultimately, the validity and utility of the NAEP program resides in its capacity for getting useful information into the hands of the right people, and in a fashion that minimizes the potential for inaccurate inferences from the data. The vehicles for disseminating information include press releases, a variety of authored reports (now available as both print and Web documents), data tabulations, and data sets that can be manipulated for further analyses. Technical documentation is another component of information dissemination, which serves to facilitate appropriate secondary analysis and interpretation of NAEP findings.

The kinds of issues with which NAEP has grappled in trying to produce valid and useful information include:

- How to provide the public with an understanding of the substance of student achievement that goes beyond a numeric score
- How to convey the real world meaning of points on the NAEP scale, and the practical significance of performance changes of different magnitudes
- How to convey the statistical error associated with NAEP estimates, and the kinds of inferences audiences should or should not make from NAEP results given this error
- How to help consumers reach appropriate conclusions about education and student achievement, given such (more or less) evident constraints as the fact that any one test is an imperfect measure of student achievement, and different tests will sometimes suggest different conclusions
- How to use information about associations between NAEP achievement scores and NAEP background questions to inform the discussion of what works in education without encouraging inappropriate causal inference or, conversely, discounting the information value of the results

A substantial amount of research has already been done on some of these questions, particularly how to add meaning to the NAEP scale and how to accurately convey complex statistical data. However, there is little belief that fully satisfactory solutions have yet been identified in any of these areas. In particular, there is considerable evidence that the primary audiences for NAEP results lack both the skills and the motivation to extract nuanced messages from generalized displays of NAEP data.

Audiences clearly do better if they are given brief, focused interpretations of NAEP results. This can be done verbally, but Wainer and others have also demonstrated that it is possible to construct focused data displays that lead readers toward desired conclusions.[1] The problem arises in that NAEP data *are* nuanced, and it is rare to find single, clear answers to the questions that people care about that are also technically defensible. In fact, even the act of framing the questions may have policy as well as technical implications. For example, NAEP has faced such policy issues in the area of state comparisons, as well as in creating definitions of adequate student performance. As a statistical agency, the National Center for Education Statistics' (NCES) stance has generally been to favor more neutral data displays, both because of the wide range of potential questions that can be addressed using NAEP findings and the desire to stay within the strict limits of the data.

The complexity of providing the public with clear and useful information is exacerbated when NAEP is viewed as one of several data sources addressing the same fundamental questions about student achievement. Increasingly, we are living in a climate where this is the case—national trends in student achievement are gauged by the Third International Mathematics and Science Study (TIMSS) as well as NAEP, and state progress is monitored by state assessments as well as NAEP. Variations across data sources underscore the ambiguity that arises in efforts to measure student achievement and, particularly, to use measures of student achievement to gauge the adequacy of our educational system. Furthermore, discrepancies across data sources, which are currently only a source of annoying complexity, will assume far greater importance if—as is currently proposed—NAEP becomes the basis for confirming or disconfirming results posted by states on their own assessments.

Clearly, if this policy goes forward, a compelling priority for NAEP validity research must be a focus on the validity of NAEP-based conclusions regarding states' progress, both in terms of overall student achievement and in closing the gap between advantaged and disadvantaged students. Such research would have to go far beyond any narrow definition of "reporting," but would ultimately have to include consideration of how NAEP findings are presented to the public, and who is given the job of deciding whether NAEP results are in fact confirmatory, not confirmatory, or neither. Here, as elsewhere, the complexity of the communication problem is inversely related to the strength of the conclusions that report writers are willing to draw.

If those charged with determining state progress are willing to sift through the available evidence and pronounce states' progress as adequate or inadequate, then it becomes relatively straightforward to frame a report to convey these findings. If, however, the goal is to lay out detailed evidence from which others can draw a variety of conclusions and evaluate the relative strengths of these conclusions, then the reporting problem is substantially more complex.

---

[1] See, for example, Wainer, H. (1997). Improving tabular display, with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics, 22*, 1-30.

In an earlier paper for the NAEP Validity Studies Panel, Richard Jaeger laid out a set of dimensions for exploring the success of NAEP reporting practices.[2] He identified three primary research questions:

1. In what form should NAEP results be reported? (That is, how should students' collective performances on the NAEP assessments be summarized?)
2. How should NAEP results be displayed?
3. How should NAEP results be disseminated?

Jaeger also felt that these research questions had to be answered with regard to specific audiences for NAEP, and he identified nine distinct audiences: the federal executive branch, Congressional staff members, state executive branches, state legislatures, district-level administrators and professional staff, school principals and teachers, the general public, members of the press, and educational research personnel.

In setting priorities within this very broad area of research, Jaeger suggested that the first priority be given to the investigation of reporting through the public media. His rationale for this choice was that "members of the general public and many policymakers receive their information on NAEP results either primarily or solely through the public media."[3] The goal of the investigation would be to identify those factors under the control of the NAEP program that would best support accurate and useful reporting by the media.

Other priority research areas that Jaeger identified were:

- Making NAEP reporting more understandable and useful to school curriculum and instruction personnel
- Reporting to the public
- Further research with state education personnel (this latter being the area in which most research on NAEP reporting has been done to date)

The problem with recommending a broad and open-ended research agenda for reporting at the present time is that there is little evidence to suggest that it is possible to make rapid gains in the general interpretability of NAEP reports, especially given the broad and varied purposes for NAEP reporting and the various constraints on data interpretation enumerated above. This is another argument for focusing any validity research related to reporting on a few clearly defined reporting problems—currently, the most critical must be the approach NAEP should take toward confirming states' progress.

The first step is primarily psychometric and statistical and, as such, reaches beyond a narrow definition of reporting. Specifically, NAEP must continue the process, which has already begun in an informal way, of analyzing the attributes of NAEP scores and assessment scores from various states and determining the kinds of confirmatory or not confirmatory evidence that NAEP could potentially provide. This task is made more challenging by the relatively small sizes of state achievement gains expected, both overall and relative to closing the gap between advantaged and disadvantaged students. This analysis could potentially lead to recommendations for modifying the NAEP instruments

---

[2] Jaeger, Richard. (1998). *Reporting the results of the National Assessment of Educational Progress*. Palo Alto, CA: American Institutes for Research.

[3] Ibid., p. 31.

(e.g., improving measurement accuracy in the lower region of the scale) or data collection procedures (e.g., increasing sample sizes to decrease sampling error) in order to enhance NAEP's suitability to this purpose.

Once the limitations on NAEP/state comparisons are fully understood and NAEP has been fine tuned as necessary to enhance these comparisons, a more classical reporting problem arises in determining how best to convey the comparisons. A related issue, which is more a matter of policy than research, concerns the role that NCES and the National Assessment Governing Board (NAGB) might play in actually carrying out the comparisons. That is, the reporting problem is somewhat different if NAEP establishes the criteria for states' success or simply makes data available to some other entity which is then responsible for passing judgment. Once there is clear consensus on what NAEP ought to report, focused cognitive lab studies could be carried out to determine which data presentations are most successful to conveying this information to key audiences. (This approach would be in keeping with Jaeger's observation that individuals typically work alone to extract meaning from reports, and this meaning-making experience is not well captured by group-focused research.)

No matter who ends up being officially responsible for evaluating states' claims regarding progress, the higher stakes environment is likely to shape the information demands that NAEP itself must face from policymakers, the press, and the public. NAEP should establish an ongoing monitoring effort to track the kinds of questions people seek to answer in the new policy environment and the extent to which NAEP data are deemed useful and/or used appropriately. That is, as conditions change NAEP should not assume that the current reporting choices remain the most appropriate.

Finally, the pressures for frequent and rapid reporting that are attendant upon the proposed augmentation of NAEP's role are likely to hasten the move toward providing more web-based data for tailored reporting and secondary analysis by end-users. Most research on reporting has focused on improving the accuracy of inferences from *authored* reports. Now that we are building much greater technical capacity for secondary users to analyze NAEP data directly (particularly with the on-line data tool), concerns also arise as to the best ways to insure accurate inferences from these home-grown analyses. Possible solutions range from offering more information on appropriate and inappropriate analyses to actually building in software constraints that preclude certain kinds of analyses. This is another area in which research could be proposed, and it would be particularly interesting to determine the extent to which education reporters begin to make use of these web-based data and whether the inferences that they draw are more or less defensible than the ones that they extract from traditional printed reports.

Chapter 7. Estimating Trends from NAEP Scores: Rationale and Research Directions

Subcommittee: *David Grissmer*
Albert E. Beaton
Larry Hedges

The Need for Trend Measures

The main focus of attention when data from any of the three NAEP assessments (long term, main and state) are released is whether the new data show a "significant" improvement from previous tests. We assume here that a statistical test is required to determine if a "significant" gain has been made in achievement scores. The method currently used to answer this question is to determine whether the difference in the means of the recent and past scores are statistically significant at the 5-percent level using the estimated standard errors of the scores and an "appropriate multiple comparison procedure." The standard errors take into account variability due to student and question sampling only. The multiple comparison procedure makes the test more stringent and essentially tries to eliminate those statistically significant results that would result from purely random draws. When there is more than one previous test, a comparison is done between the current score and each previous test score.

An alternative or supplemental method to determine whether significant gains are present is to use trend estimation over the entire or a partial range of scores. In some previous years, NAEP documentation provided estimates of trends in long-term scores using both a linear and non-linear (quadratic term) specification. However, the argument here is that it may be useful to go beyond simply providing trend estimates and make the trend estimates a major focus of the interpretation of scores for the public and press.[1] There are several problems with the current two-way comparisons that might be addressed through adding trend estimation:

- The standard errors reflect only a portion of the error for each test, and thus the standard statistical tests probably overstate what is actually a significant change from one test to the next.[2]

---

[1] Both linear and quadratic estimation might be useful in a policy context. Linear trends would provide a kind of benchmark when comparing the annual gain since the last test with the long-term historical average trend. A quadratic estimation would provide information about the change in trend, and whether the current gain matches gains in recent years.

[2] ETS has also found that estimated standard errors are too small using a double jackknifing procedure. (private communication with Eugene Johnson.)

- Since each comparison includes the most recent score, a large error in that score will significantly affect all comparisons. Essentially, the current analysis relies too heavily on the most recent data point.
- The two-way comparisons cannot readily be interpreted to determine whether a consistent pattern of improvement is present that is statistically significant. For instance, non-significance of most of the two-way tests could hide significant trends when all tests are analyzed together, and many two-way tests could be significant without a significant trend being present.
- The current procedure does not account for the length of time between the current test and previous tests. If gradual improvement in scores is occurring, we would expect a higher probability of statistically significant results the longer the time period between the tests being compared. Yet the current procedures give equal attention to comparisons between tests two years apart with those four or eight years apart.

Trends implicitly can reflect all sources of random variation in a test score, not just student and question sampling error, since the actual variation in data points serves as the basis for statistical tests. Trends also implicitly equally weight each data point, eliminating the undue weight given to the current score in current analysis. Additionally, the statistics generated by trend analysis incorporate the consistency of improvement over time as well as the magnitude. Finally, trends also are able to take account of the time between tests and estimate annualized gains.

Even with accurate trend estimates, a second problem of interpretation arises. What are the trends measuring, and what should they measure? The purpose of the NAEP scores is to monitor "real" progress in achievement. Certainly, we do not want trends to reflect the influence of changing participation rates and exclusion rates, nor the influence of such factors as changes in the timing of the tests, or changes in the age distribution within grade. However, whether we want trends to reflect changing demographics is more problematical. Changes in demographics (as shown below) have influenced NAEP trends. On the one hand, a trend that adjusts for demographic changes may be a better measure of school improvement. On the other hand, a trend that reflects scores of actual students, even through the characteristics of those students has changed, is also useful in that higher achievement for all students is a goal.

The purpose of this paper is 1) to explore some issues with the current procedure of using two-way comparisons as the principal statistical method to determine if "real" progress has been made in achievement, 2) to explore whether adding trend analysis could provide improved measures of "real" progress, and 3) to outline a research agenda that would be needed to make the necessary trend estimates.

In this paper, the question of whether we should proceed with making trend estimates is separated from the question of whether we should publish such estimates with full NCES backing. Estimating such trends provides very useful information that helps interpret NAEP scores, and we should probably move ahead with these estimates. Whether trends should become a central focus of interpretation of NAEP scores in NCES publications depends partly on political interpretations about the purpose of NAEP, and partly on analytical judgment about what approach conveys the most accurate and useful information in support of its designated purpose. Time and resources do not allow this paper to be definitive, but hopefully it will provide enough information to serve as a basis for further discussion and decisions about proceeding with a research agenda.

The next section describes an approach that helps to determine when two-way comparisons are appropriate, and when trends might be appropriate. This approach requires estimates for actual trends, standard errors, and other sources of noise. These estimates are developed using the long-term NAEP data and the state NAEP data, and the results are interpreted to suggest that two-way comparisons can provide misleading results for both national and state results. Finally, this paper concludes with a research agenda needed to further explore the issue of trend estimates.

Approach

Two-way comparisons work best when the "expected signal" or expected annual rate of improvement in scores is much larger than the "noise" (standard error + other sources of noise) in the system. In this case, the comparison of scores with intervals of one or two years apart would be expected to produce statistically significant results if expected progress had been made. Since statistical significance is the current critical test for whether progress is being made, the two-way comparison would provide a good measure.

Two-way comparisons work less well when the signal is approximately equal to or less than the noise. Two cases present themselves here. If the standard error is much larger than other sources of noise, the two-way statistical tests over short intervals (two to four years) will usually not produce statistical significance, even when the expected annual progress is being made. So two-way comparisons can be misleading in this case since no statistical significance would indicate no "real" progress, when in fact the expected progress is being made. It is only in the longer term that such progress can be identified, and trends can serve as a better measure of whether steady progress is being made.

In the alternate case in which sources of noise outside the standard error can be as large or larger than the standard error, a different scenario unfolds. Here the standard two-way comparison based on the standard error will produce many false positives— a statistical significance when no real progress is being made. In this case, trend estimates would be better because they incorporate the other sources of noise in their statistical tests, and thus provide a better measure of whether progress is being made.

In a simple model, the choice of trends over two-way comparisons depends on four parameters: the signal (S), the standard error (SE), the magnitude of sources of noise not included in the standard error (SN), and the interval between tests (I). We only have some control over SE and I in the NAEP design. The sample size partially determines SE, and the interval between tests is set by policy, but usually without consideration of the interconnection between the four parameters. The signal is determined by the magnitude of "policy significant" score increases over time. We have very limited control over SN beyond a certain point. We try to adopt procedures that minimize SN, but reach limits beyond which further reduction is not possible or too costly.[3]

---

[3] SN can include many factors such as changes in the timing of administration of tests, changes in administrative procedures, changes in participation and exclusion rates, etc.

So it is necessary to look at current values for S, SE, N and I to determine if two-way comparisons are conveying accurate information. The long-term trends and state NAEP need separate analyses since the parameters take on somewhat different values.

Analysis of Long-Term Trend Data

The question of what constitutes a policy significant gain in NAEP scores cannot be separated from empirical estimates of actual gains, or what is actually possible. Political goals have often been set that are completely unrealistic in terms of empirical experience. Ultimately, a policy significant gain in achievement is determined by the political and economic environment, and it is sometimes related to closing a "gap" in scores, either internationally or between racial/ethnic groups in the U.S. But, the missing parameter in such calculations is usually how long a period might be required to close such gaps. These relevant gaps can be 25 or more percentile points.

Gains of one percentile point in a decade would not be significant from any policy perspective. Gains on the order of two and a half percentile points a decade arguably begin to reach a threshold that one might consider should be detectable by NAEP. Clearly gains of greater than five percentile points a decade should be detectable. For these calculations, we somewhat arbitrarily assume that gains of .25–.50 percentile points a year are in the range of scores that NAEP should detect in their interval between testing. Are these gains empirically possible?

Table 7.1 shows the results of fitting simple trends to the long-term NAEP assessments from the earliest test (1971 for reading and 1973 for mathematics) through 1999. The trend variable reflects the number of years between tests, thus implicitly assuming a constant annual rate of improvement. Two estimates are provided: the unadjusted scores use the actual scores, while the adjusted scores reflect scores if demographic characteristics had remained constant over the test period.[4]

The unadjusted results show positive trends for each of the tests, with statistically significant positive trends for five of the six tests. The range of annual improvement is .07 to .64 percentile points. Mathematics scores show much higher rates of improvement than reading scores. Adjusting for demographic changes makes all trend coefficients larger and more highly statistically significant. Annual improvement estimates range from .16 to .70 percentile points. These results would suggest that demographic trends have lowered actual scores, and account for a non-significant portion of the variance over time—especially for reading scores.

---

[4] The trend scores are converted from scale points to standard deviation units by dividing by the earliest standard deviation, and then to percentile units by multiplying by 34.5. The adjusted trend scores are estimated by assuming the demographic mix for the earliest test, and estimating each subsequent score by using the racial/ethnic scores for each year.

Table 7.1–
Estimates of the Rate of Uniform Gain (in percentile points) for the Long Term NAEP Data from 1971–1973 to 1999

	Unadjusted			Adjusted for constant demographics		
Test	**Trend coefficient**	**t-value**	**R-squared**	**Trend coefficient**	**t-value**	**R-squared**
Reading- Age 9	0.07	0.97	**.11**	0.16	2.17	**.37**
Reading- Age 13	0.12	3.15	**.55**	0.20	5.79	**.81**
Reading- Age 17	0.13	2.32	**.40**	0.25	4.04	**.67**
Mathematics- Age 9	0.64	6.95	**.87**	0.70	7.78	**.90**
Mathematics- Age 13	0.44	8.24	**.91**	0.52	9.47	**.93**
Mathematics- Age 17	0.28	2.92	**.55**	0.38	4.00	**.70**

The R-squared statistics show that a simple trend accounts for a large portion of the variance in mathematics scores, but also a significant portion of the reading scores variance with the exception of the 9-year old scores. Since the 9-year old reading score has the lowest trend, it would be expected that noise accounts for more of the variance than for the other estimates. The results would suggest that mathematics gains are clearly significant from our definition of policy significant gain, while reading gains are below, or border on, policy significant gains.

The next question is how these gains compare to the standard errors. The standard errors for these national scores range from approximately .8 to 1.3 percentile points. So for national mathematics scores, the signal—the annual gain rate—is somewhat less than the standard errors, while for reading, the signal is much less than the standard errors.

Since the racial/ethnic gaps in scores are a major policy issue in this country, designing samples and intervals to detect policy significant gains by racial/ethnic group is important. Figure 7.1 shows the trend estimates by racial/ethnic group. All nine mathematics trends by racial/ethnic group are statistically significant at better than the 1-percent level. For reading, eight of nine trends are statistically significant at the 5-percent level. For practically all age groups and both subjects, black gains are the largest, followed by Hispanic gains, with white gains being the smallest. Mathematics gains exceed reading gains for each age and racial/ethnic group. There are few differences by age within each subject and racial/ethnic group. All comparisons except reading gains for white students show gains that could be classified as policy significant gains.

For the racial/ethnic scores, the standard errors for white students can be as large as 1.5 percentile points, while the standard errors can be as large as 2.7 percentile points for black students and they can exceed 4 percentile points for Hispanic students. For all racial/ethnic groups, the annual estimated gain is significantly less than the standard errors.

These trends suggest that a model that assumes that educational progress has occurred at a steady rate in both subjects and for all ages and racial/ethnic groups is not easily dismissed, but the gain rates differ by subject and racial/ethnic group. It also shows that annual progress is always less than, and often much less than, the standard error.

The results also imply that it would take, at best, four years, but usually well over eight years, to achieve gains that would exceed standard errors by a factor of two—roughly the 95-percent confidence test. Current sample sizes are thus too small to detect policy significant gains with two-way comparisons over intervals less than eight years for most tests, age groups, and racial/ethnic groups.

Figure 7.1–
Estimated Annual Gains in Percentile Points for Long Term NAEP Scores

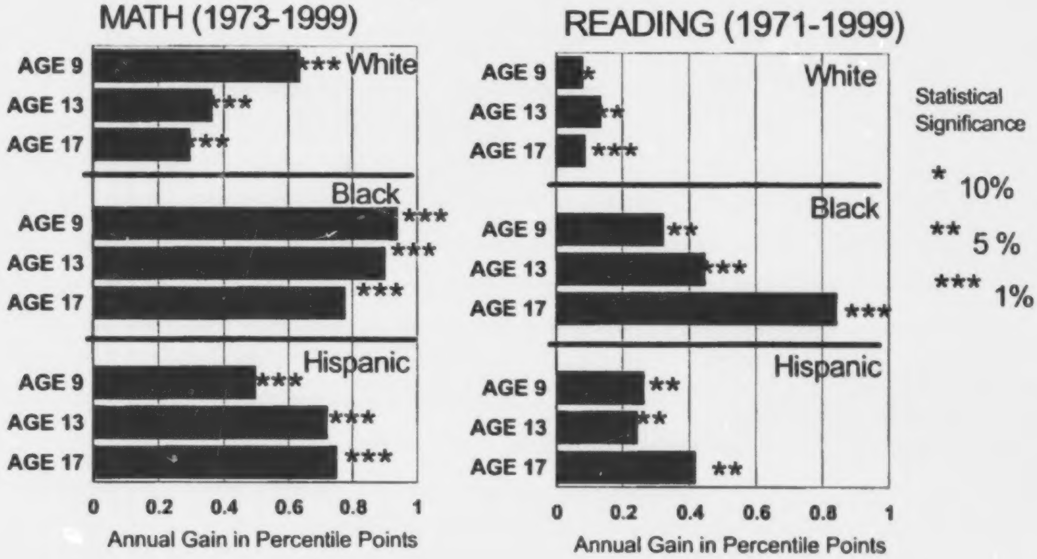


Table 7.2 illustrates this point by showing the number of statistically significant differences measured between 1999 long-term scores and previous years. This comparison includes nine comparisons per year (three racial/ethnic groups and three age groups). There is only one statistically significant difference comparing 1999 to 1996 or 1994. Four of 27 comparisons show statistical significance for the 1992 scores, and increasing proportions show significance for earlier years. These results suggest that the annual rate of improvement in scores cannot be separated from the noise over periods shorter than seven years for mathematics scores, and much longer for the slower improvement in reading. The NAEP sample would need to be significantly increased to make two-way comparisons regularly detect ongoing improvement over the current two-year interval.

Table 7.2–
Number of Statistically Significant Differences from 1999 Scores for Long Term NAEP Scores

Comparison Year with 1999	Number of Statistically Significant Differences		
	Reading	Mathematics	Science
1996	0	0	0
1994	1	0	0
1992	0	3	1
1990	1	4	2
1986 (mathematics, science) or 1988 (reading)	1	6	4
1982 (mathematics, science) or 1984(reading)	1	9	8
1977-8 (mathematics, science) or 1980 (reading)	7	9	9

Table 7.2 also implies that for national scores sources of noise other than those incorporated into standard errors may not be a major factor. If SN>SE, then we would expect to see statistically significant differences even over short time intervals. While SN could still increase the number of statistically significant differences in Table 7.2, the fact that it does not do so over the five-year period from 1999–1994 places some limit on its magnitude.

Analysis of State Trend Data

The national NAEP results suggest that conditions may be present that make two-way statistical tests problematic because the standard error is too large to detect policy significant changes in scores over the two- to four-year NAEP intervals. State test results show a somewhat different pattern.

Grissmer and his colleagues estimated state trends in NAEP score improvement for the 1990–1996 tests.[5] Recent updates using the 1998 and 2000 scores show little difference in the range of estimates. The results show almost no state making statistically significant reading gains, but many states making significant mathematics gains that range from zero to two percentile points a year. These estimates include some adjustment for changes in demographics, participation, and exclusion rates. Compared to national gains, individual states can make gains much larger than the nation as a whole.

The standard errors for state tests range approximately from one to two percentile points. So, the annual improvement for states making the most rapid improvement can be of the same magnitude as the standard error, but is less for states with improvement rates below one percentile point a year (about one-half the states). In the best-case scenario for state scores, two-way comparisons could detect the gains occurring in the high-gaining states over a four-year interval, but more commonly it would take six or more years to detect statistically significant two-way comparisons. Since state tests have been given over two-, four-, and six-year intervals, we would expect to see almost no statistically significant gains for reading among states, and some statistically significant gains in states with the highest gains over four- or six-year intervals.

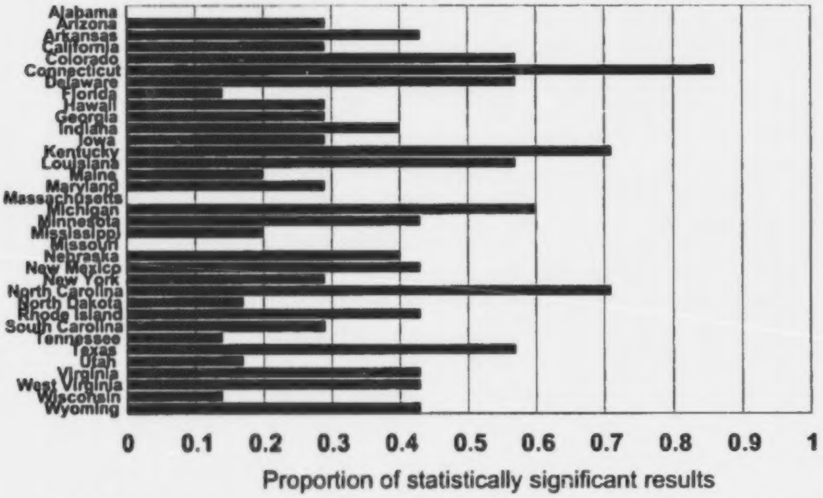
---

[5] Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. Santa Monica, CA: RAND.

However, unlike the national tests, the state tests show large numbers of statistically significant results over two-, four-, and six-year intervals. Figure 7.2 shows the proportion of statistically significant results in the state score comparisons from previous tests in the period from 1990 to 1998. There were 230 comparisons made and 83 comparisons resulted in statistical significance, i.e., the chances of a state getting a * in any single comparison with a previous test was greater than one in three. Only three states failed to get at least one statistically significant gain. If there were no systematic trends in the data, and only student and question sampling sources of error, the author maintains that we would expect less than 12 significant results, and maybe close to none based on the multiple comparison tests. So the large number of significant results suggest either that trends are present, or that there are significant sources of error (SN) outside of the measured standard errors. But in the absence of trends, these results would suggest that SN would be a significant factor in determining state score significance.

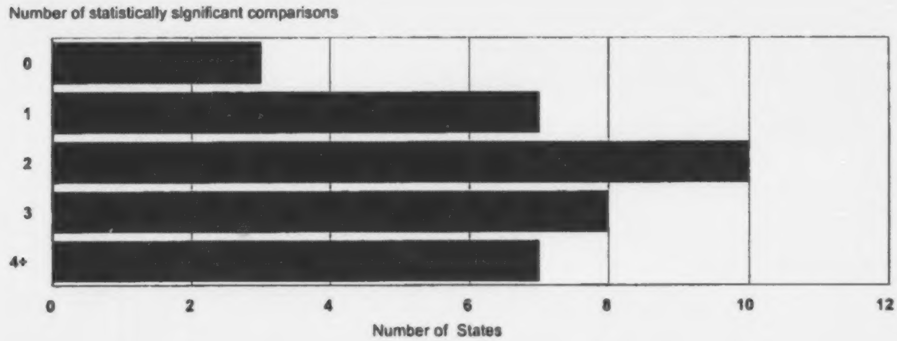
Whether the source of so many significant results is trends or sources of noise can be explored by looking at the distribution of results. If a lot of states had strong trends, then we would expect those states to account for a high proportion of significant two-way comparisons. If noise were the source of the significance, then the author maintains we would expect a distribution with few states having many significant results and many states with a small number of significant results. Figure 7.3 shows the distribution of results by states. Each state averaged 6.6 comparisons (some had seven and some only five), and the distribution shows only seven states had four or more statistically significant comparisons. The large number of states having from zero to two significant comparisons hints that many of the results designated as statistically significant at the state level may be due to noise (N) rather than significant trends. It is also the case that statistical significance was as likely at a two-year interval as at a four-year interval.

Figure 7.2–
Number of Statistical Differences in Two-way Comparisons for State NAEP Scores (1990–1998)



This analysis would suggest that for state scores, unlike national scores, N can be of the order, or larger than, the standard error. There are several sources of noise that might be expected to be larger at the state level than the national level. These sources can include changing demographics, participation rates, exclusion rates, timing of administration and changing age distributions. For instance, demographics in many states like Texas and California would be expected to change more rapidly than national demographics. The timing of test taking by state would likely be less consistent over the years than the national tests. Participation and exclusion rates also have more variance at the state level.

Figure 7.3–
Distribution of Statistically Significant Results for States



For state scores, the argument for trends is different than for national scores because noise looks to be a much more significant factor. For states, the reason to estimate trends over two-way comparisons is that the latter give many false positives, whereas trends, since they incorporate all sources of noise into the statistical tests, may provide a more accurate picture.

Summary

The heart of the current policy debate is whether states and the nation can make structural reforms in education that lead to policy significant sustained improvement in educational outcomes. Being able to provide data and statistical tests at regular intervals to determine if such sustained gains are taking place is the primary purpose of NAEP. This paper argues that, given current sample sizes and intervals between tests, trends need serious consideration as an alternative or adjunct to two-way comparisons in determining whether policy significant gains are occurring for both national and state scores.

Estimating trends rather than two-way comparisons is particularly important when there are significant sources of random variability (SN) not captured by standard errors (SE), and the total variability from all sources of noise is about the same or of greater magnitude than the difference in signal in the test interval (I).

Two-way comparisons may be problematical in both national and state NAEP. In national NAEP, current sample sizes practically guarantee that two-way comparisons will be insignificant from the previous test administered either two, four, or six years ago, and thus not provide useful information about whether progress is occurring in shorter than eight-year intervals. In the state case, two-way comparisons may overstate the significance of results due to a higher level of noise in state results than national results. In either case, trends could possibly provide a more accurate and transparent picture of the progress in scores.

In the national case, trends show statistically significant progress for practically all age groups, racial/ethnic groups and subjects, whereas two-way comparisons using the most recent test nearly always show insignificant gains. The focus here might be whether the latest score showed a similar trend as in the past, rather than whether it is statistically significant from the previous score. For the state scores, trends incorporate all sources of noise, and thus it seems essential to separate states making sustained gains from those large numbers of states showing significant gains from the most recent test.

However, determining whether trends can become a viable alternative or an adjunct to two-way comparisons needs considerably more refined analysis than the more suggestive analysis given here. In addition, the factors that might bias trend estimates, making them problematic, needs to be determined. We turn to that research agenda now.

Research Issues

It is becoming increasingly important for two-way comparisons to account for changes in scores that might be caused by changes in exclusion and participation rates, timing of test administration, and in some scenarios, changes in demographics. State scores especially will require adjustments in order to have credibility. Any trend estimation will also need to incorporate similar adjustments. So in either case, the research agenda needs to include the development and assessment of alternative ways of doing such adjustments.

However, trends may have two advantages in this regard over two-way comparisons. First, two-way comparisons need to develop adjustments outside the estimation process, whereas trend estimation has the option of using this technique or estimating such adjustments directly in the trend estimation process. Essentially, the two-way comparison methodology will develop an adjusted score corrected for changes in a certain factor, such as exclusion rates, before making statistical tests. These adjusted scores can also be used in trends. Another option is to include these factors on the right-hand side of the equation and use the entire set of data to determine the appropriate adjustment. Research is needed to develop and compare such methods. The second possible advantage is that trends will be less sensitive to test-to-test changes in these factors than will two-way comparisons, and thus some of the political issues associated with such changes may be minimized.

The long term NAEP, the main assessment, and state NAEP all present somewhat different challenges to developing adjustments and to trend estimation. Trends inherently require at least three data points to generate statistical estimates with standard error estimates, bur realistically trends need more than three points. The long term NAEP presents the easiest case since there is a much longer time period, and at least nine tests have been given per subject and age group from 1969 to 1999. The main assessment can provide estimates from 1986 at best, but comparable data probably exists from 1990, thus providing only five tests. State NAEP has four tests for eighth-grade mathematics over 10 years, but three or fewer for the remaining tests. Assuming the same rate of improvement across grades, the number of data points can be increased by pooling across grades in state and national tests. However, additional data will continue to add data points to all three NAEP tests making improved trend estimation possible.

There are several variables that can introduce systematic bias in achievement scores that require adjustment if "real" gains are to be separated from spurious gains. Demographic changes occur systematically over time, as have increases in exclusion rates and participation rates. Other possible sources of bias include the changing age distribution of the test takers and perhaps the timing of test administrations. One-time changes in procedures, that may have caused the downward bias in 1994 reading scores, also needs to be accounted for.

Developing a methodology for estimating trends would require using methods that attempt to account for these sources of bias in scores. Research is needed that would assess whether and how to adjust for the most serious sources of bias: changing participation rates and exclusion rates, demographic changes, changes in timing of administration, and changes in age distributions.

There are two broad methods to approach these estimates. The first is to make estimates of the magnitude of bias from statistical assumptions (e.g., the analysis of several states' data in the 1998 reading tests by Don McLaughlin). The second method is to derive estimates through specification of a model for achievement scores that includes variables for sources of bias. Such an approach is used by Grissmer and his colleagues to correct state scores for changes in participation rate, exclusion rates and demographic changes.[6]

Ideally, research would use both individual-level and more aggregate-level scores pooled across all time periods to estimate score trends and adjusted score trends for each of the three NAEP samples. Table 7.1 has such an example using aggregate data of estimated trends, both adjusted and not adjusted for demographic changes. The value of estimates using individual-level data is that most adjustments such as changes in age distribution or timing of administration could be estimated more accurately than at the aggregate level. The research should also estimate trends and adjusted trends using scores that have been adjusted using statistical models outside the trend estimation, such as Don McLaughlin's full population score estimates.

---

[6] Ibid.

The research to be conducted would involve separate consideration and estimation using the long term NAEP, the main assessment, and state NAEP tests. Each would need separate analyses to develop adjustments and to assess the best methodology for trends.

A simplistic example of individual-level estimation would be:

$$\text{SCORE}_{ij} = a + \sum_k b_k x_{ijk} + tz_{ij}$$

where Score $_{ij}$ is the score of the ith student in year j, x $_{ijk}$ is the kth adjustment factor for student ij and z $_{ij}$ is the year of the test for student ij.

The coefficients to be estimated are a, b_k and t. The x's can include variables such as the age of the student in months, the time of the year for administration, race/ethnicity, measures of parental education, or free lunch participation. A variety of trend estimates, t, can be made by including or excluding all or some of the x's in the estimation: eliminating all would produce an unadjusted trend, adding a racial/ethnic variable would produce a trend adjusted for constant demographics, and adding an age variable would produce a trend assuming a constant age distribution. Although the equation is stated as linear, nothing would preclude including non-linear terms if such effects are present and significant. While simple in concept, the actual estimates would be more complicated due to weights, missing data and many other considerations.

Importance, Implications and Costs

As more NAEP data accumulate, stretching over longer time periods, trend analysis may be needed to either replace or supplement the current methodology for determining whether significant progress is being made in achievement. A plausible case has been presented that two-way comparisons for both national and state scores do not accurately convey whether policy relevant progress in achievement is being made. In national scores samples sizes are insufficient to detect policy relevant gains over two-, four-, or even six-year periods, leaving the impression that nothing significant is happening in recent years. But, long-term trends indicate that highly policy relevant and statistically significant gains *are* occurring. For state scores, the issue for two-way comparisons is the presence of many false positives over two- and four-year intervals because of high levels of noise. These two-way comparisons show progress where none is occurring. Trends are needed here in order to incorporate the full uncertainty into the statistical determination of whether policy relevant gains are occurring.

From a policy perspective it seems important to shift focus from measures that either cannot detect policy significant gains in the short run, or that contain many false positives, to ones that measure whether sustained gains are occurring over longer time periods. We know that educational change is slow, not dramatic, because significant structural and organizational changes are needed to improve achievement. The key question emerging is whether we are seeing a pattern of consistent changes in scores over time due to adding resources or through structural, systemic reform efforts. It is slow steady progress that needs to be separated from the noise, and trends are probably the simplest method of doing this.

The value in carrying out such research on trends is not only to assess whether trend estimation is feasible and can draw a more accurate picture of NAEP performance, but it will allow NCES to be more proactive in addressing bias issues in NAEP data. Our current approach has been reactive in the sense that we wait until a problem surfaces in the scores, and then try in a brief time period to address the issue. Conversely, the research project suggested here would deal proactively with the major sources of bias and incorporate them into the methodology, and hopefully avert some future crisis.

Moving toward trend estimation would require a significant research project based on combining the individual data sets within each of the three assessments (long term, main and state NAEP) and carrying out a variety of estimations. Some of this work has been started in various places, but a significant amount of work is probably needed. Much analysis will be required to test various methods and to determine which methods correct more accurately for each source of bias. A rough estimate for such an effort would be $750,000 to $5,000,000. This estimate would include review and publication of the results that would be crucial for credibility. Once methods are developed, the operational costs would be approximately $100,000–$200,000 for each set of tests given in a year.

Chapter 8. Synthesis: An Agenda for NAEP Validity Research

Since its inception over 30 years ago, the National Assessment of Educational Progress (NAEP) has been a pillar of methodological rigor in achievement testing. Throughout its history, NAEP has not shied from addressing difficult problems in reporting valid achievement statistics for the nation. Despite massive shifts in methods and scope, such as the introduction of psychometric scaling in 1984 and the extension to state-by-state reporting in 1990, NAEP has held to its mission of measuring the progress of our nation's educational system in equipping our young people with the skills needed in the modern world.

The NAEP Validity Studies (NVS) Panel continues that mission with this report, laying out an agenda for validity research in the context of NAEP's new role in American education: to support states in their assessments of their schools' annual yearly progress (AYP).

The preceding chapters have described important validity studies addressing questions of what is tested, how it is tested, who is tested, and how the results are analyzed and used, with special emphasis on trends. This chapter lays these studies side by side and provides indicators of priorities. The NVS Panel met in November 2001 and discussed each study area; each panel member then recommended a priority for each study, ranging from "E" for "essential" to "N" for "not worth doing."

The Study Areas

Validity, as used here, refers to the quality of NAEP reports: that they convey and do not distort the picture of elementary and secondary scholastic achievement in the United States. The contexts and processes of gathering, analyzing, and reporting data can distort inferences about scholastic achievement in many ways; research is needed to assess and, if needed, remove these "threats to validity." The NVS Panel initially partitioned the threats to validity into five aspects: 1) what is tested, 2) how skills are measured, 3) who is tested, 4) how the data are analyzed, and 5) how the results are used. However, many threats to validity are operative primarily in the context of measuring and reporting trends over time, so trend reporting is included as a sixth, separate category of validity research.

What Is Tested?

When NAEP reports that 25 percent of the fourth-grade public school students in a state are reading at a proficient level, what does that mean? Is the activity that NAEP calls proficient reading the same thing that state school administrators, local teachers, or parents call proficient reading? If NAEP's definition of a subject area is different from that of a key constituency, then the validity of NAEP's reports can be challenged. Validity research is needed that assesses the convergence of NAEP and other definitions of the domains of achievement on which NAEP reports.

How Are Skills Measured?

Again, when NAEP reports that 25 percent of the fourth-grade public school students in a state are reading at a proficient level, what does that mean? Although NAEP, state school administrators, local teachers, and parents may all agree that what they want reported is the same "reading proficiency," they may not agree that NAEP is testing that proficiency. Some people may feel that NAEP requires too much writing on the reading test, or too little, or that NAEP requires too much reading on the mathematics test, or too little. A myriad of factors can affect test scores, and each is potentially a threat to the meaningfulness of NAEP to some audiences. Validity research is needed to assess the impact of the most important dimensions of "contamination." In particular, in the context of the movement to include all students in testing, accommodations, or non-standard administrations, are provided for some children to remove barriers that would otherwise interfere with their ability to show what they know. If these accommodations change the constructs tested, the validity of the resulting scores can be challenged.

Who Is Tested?

Once more, when NAEP reports that 25 percent of the fourth-grade public school students in a state are reading at a proficient level, what does that mean? NAEP administers its assessment to a sample of students in public schools in a state and uses the responses of those students to estimate proficiency in the state. The students invited to participate in the assessment are selected at random from a list of fourth-grade students in a randomly selected sample of schools drawn from a list of public schools in the state. Therefore, generalization to the population is potentially valid. However, what about schools that are not on the list, schools that refuse to participate, students who are not on the list, and students who do not participate? If a segment of the student population is missing from the sampling frame or under-represented in the sampling frame, then NAEP's proficiency reports can be challenged as not fairly representing the state of educational achievement.

How Are the Data Analyzed?

The psychometric sophistication of NAEP's method for estimating population performance is unparalleled, but it takes substantial time to implement the processing and analysis of each assessment's data. Existing pressures on NAEP to present results sooner will only increase with NAEP's new role in evaluating and corroborating state assessment results, and procedures for increasing the speed of analysis have been proposed. The serious validity questions that arise are whether and to what extent analytical shortcuts preserve the basic message of the data. If the proposed shortcuts produce results that are different from the results produced by the current analytical method, implementing them will undermine the credibility of NAEP.

How Are the Results Used?

NAEP operates in a political context in which evidence of educational achievement and educational progress is valuable. However, even with rigorous testing, precise analysis, and clear reporting, results are open to interpretation and misinterpretation. To the extent that NAEP can anticipate misinterpretation and preempt it through clarity, the credibility and validity of NAEP can be enhanced. In the context of NAEP's new role, the most important use of the results is likely to be in evaluating state assessment gains. Therefore, it is in this area that NAEP should concentrate its efforts to monitor public understanding of reports and to periodically update perspectives on the requisite data to report and the best models for reporting them.

How Are Trends Measured?

Although people recognize that achievement varies across the demographic spectrum and that many young people lack important skills, they maintain the hope and expectation that achievement will improve. Consequently, NAEP's most important reports focus on changes over time. Measuring this change is challenging, however, because the context of NAEP changes at the same time that achievement changes—for example, there are frequent calls for modernizing the content frameworks; for adding new types of items; for accommodating test administrations for students with disabilities; and for modifying sampling, scheduling, and administration. Although each factor may be constant across a single assessment so that comparisons (for example, between states) are valid, the validity of trends from one assessment to the next is threatened. Controlled research on the effects of procedural changes is needed to ensure the validity of the trends from year to year.

Studies or study areas that address these six categories of validity issues are described in chapters 2 though 7. To provide a foundation for synthesizing them into a research agenda and setting priorities, we have summarized the research as a set of 22 "studies," as listed in Table 8.1.[1] The entries in Table 8.1 represent abstractions of the particular studies described in the preceding chapters. Although the studies are briefly described in presenting the results of the priority ratings in this chapter, readers should refer to the corresponding chapters for a thorough understanding of what is being rated. NVS panel members participating in the priority rating were familiar with the papers from discussions at two previous meetings; as preparation for discussing and rating priorities, each study was briefly described by one of the authors.

---

[1] During the priority setting session, there was general consensus that the studies could be rated on importance; however, in two cases, the panel modified the original list of studies to eliminate rating problems. Entries 2, 3, and 4 in table 8.1 were originally a single entry, but most NVS panel members felt that the importance of the study depended on the nature of the question addressed, so they are prioritized separately. Also, entry 11 was added to reflect a consensus of NVS panel members that the combined importance of the preceding four entries is greater than is represented in the ratings of the individual components.

Table 8.1–
Summary of Major Validity Studies, by Assessment Aspect

What is being measured?

1. Alignment with state standards
2. Construct definition (a) What is being measured
3. Construct definition (b) What students do on the test
4. Construct definition (c) Comparison with curriculum/what is taught in the classroom

How is it being measured?

5. Contaminations
6. Accommodations

Who is being tested?

7. School list completeness
8. Representation of non-participating schools
9. Student list completeness
10. Representation of absent students
11. Combined studies of population bias (studies 7, 8, 9, and 10)
12. Representation of excluded SD and LEP students

Are there better ways to analyze the data?

13. Direct estimation with minimal conditioning
14. Omission of subscales from primary analysis
15. Analyses using school as the unit
16. Estimation of item domain sampling error

How can appropriate uses be maximized and inappropriate uses minimized?

17. Meaning of "confirming state results"
18. Limits on NAEP's capacity to evaluate state results
19. Evaluation of audience interpretations
20. Controls/supports for secondary analysis

What special information does valid measurement of trends require?

21. Bridge studies for changes in constructs, measurement, sample, administration
22. Estimation of multi-time-point trends

Framework for Priority Setting

The importance of validity research can be evaluated in terms of the potential harm that will result if the research is not done and it is later discovered that a hypothetical threat to validity is not merely hypothetical. Some kinds of errors, such as imperfect test reliability, are expected and accepted, but it is important that the size of the error be known so that the precision of conclusions is not overstated. Other kinds of errors, such as systematic biases in the selection of students for participation, also can be damaging if not assessed and corrected in analysis.

The importance of each kind of error depends on the use to which the results are put. For example, a change from one administration to the next might bias trend measures, but for audiences interested in comparisons at a single time (for example, between urban and rural schools in a state), biases in trend measures are inconsequential. To the extent that NAEP's results are relevant to educational policy, challenges can be expected from constituencies whose positions are compromised by those results. Therefore, for each study, the NVS panel members considered the challenges that might be made to NAEP results that would be addressed by the study. These challenges might be that NAEP is "unfair," "irrelevant," or "not credible"—unfair because results do not reflect all groups' achievements equally; irrelevant because what NAEP is measuring is not what policy is concerned with; or not credible because of any of a variety of inaccuracies.

Considering such challenges, the NVS panel members rated each study on a 5-point scale (Essential, High, Moderate, Low, or Not Needed).[2] This scale combines both the perceived seriousness of a problem should it occur and the likelihood of its occurrence. "Essential" conveys the NVS panel member's opinion that if a study of the type described is not undertaken, NAEP will surely be subject to potentially damaging criticism. "High" conveys the opinion that failing to undertake such a study would be a major gamble; "Low" conveys the opinion that NAEP could withstand criticism about the error because the error is unlikely to occur or its impact is unlikely to be severe. "Moderate," of course, conveys an opinion between "High" and "Low," and "Not Needed" conveys a concern that NAEP might be criticized for conducting such a validity study.

The ratings, we must emphasize, indicate the level of need for studies in particular areas, not the details of specific studies. The NVS panel recognizes that some of the studies described are already being carried out or are being considered by others. Moreover, the panel makes no statement here about either particular study designs or authority to execute the studies. The priority ratings reflect more general perceptions of the need for NAEP to do these studies.

Research Priorities

Before expressing individual ratings, panel members shared their opinions as representatives of state assessment programs, curriculum experts, psychometricians, and survey researchers; these opinions were supplemented when needed with information provided by NCES's Associate Commissioner responsible for NAEP and by observers representing NAGB, ETS, and other organizations concerned with the validity of NAEP. Thus, each panel member combined a variety of factors according to unique weights in arriving at a single rating. Although time was not sufficient to achieve a complete consensus, as can be seen in the summary of priority rating in Table 8.2, agreement was very substantial: 70 percent of the ratings were in the modal category. Only 6 percent of the ratings, mostly concerning the priority for studies of NAEP construct validity, were more than one category removed from the mode.

During the priority-setting process, each study or study area was discussed, and although specific design questions were not debated, important issues were brought up that provide context for the priority ratings. These issues are summarized in the following sections.

---

[2] Ratings were obtained by voice, going around the table in counterbalanced orders for different studies. Two members of the panel were not present and provided ratings later on the basis of a draft of this chapter showing the distribution of ratings of those present.

Table 8.2–
Priority Rating Frequencies

	Not Needed	Low	Moderate	High	Essential
What is being measured?					
1. Alignment with state standards		1			**13**
2. Definition (a) What is being measured		1	3	4	**6**
3. Definition (b) What students do on the test		1	3	**7**	3
4. Definition (c) Comparison with curriculum	2		**6**	2	4
How is it being measured?					
5. Contaminations			1	**8.5**	4.5
6. Accommodations	1			2.5	**10.5**
Who is being tested?					
7. School list completeness		4	**7**		1
8. Representation of non-participating schools		4	**6**	1	1
9. Student list completeness		1	**9**	1	1
10. Representation of absent students			**6**	**5**	1
11. Combined studies of population bias			1	**12**	1
12. Representing excluded SD and LEP students			2	**8**	4
Are there better ways to analyze the data?					
13. Direct estimation with minimal conditioning			3	**8**	2
14. Omission of subscales from primary analysis			Not rated		
15. Analyses using school as the unit			Not rated		
16. Estimation of item domain sampling error				6	**8**
How can appropriate uses be maximized and inappropriate uses minimized?					
17. Meaning of "confirming state results"					**14**
18. Limits on NAEP's capacity to evaluate state results					**14**
19. Evaluation of audience interpretations		1	**6.5**	4.5	2
20. Controls/supports for secondary analysis		**12**	1		
What special information does valid measurement of trends require?					
21. Bridge studies					**13**
22. Estimation of multi-time-point trends			1.5	**11**	0.5

Note: The modal category for each study is indicated in bold. Counts in rows vary slightly owing to a small number of cases in which a panel member was temporarily absent from the rating process or expressed an inability to recommend a rating. Fractional ratings represent individual expressions that the priority was between two adjacent categories, such as "Moderate" and "High." Studies labeled 14 and 15 were not rated, as discussed in the text.

Validity of What Is Tested

Within the context of Public Law 107-110 (No Child Left Behind), which is likely to call on NAEP to play a role in evaluating the results of state assessments, there is great concern that people who disagree with NAEP results in a particular state will argue that NAEP is not assessing skills in the form that the state mandates. There may be 50 different sets of standards for reading and mathematics, and there are definitely differences between NAEP and each state assessment. The panel felt that it may well fall upon NAEP to measure the alignment of its standards with each state's standards, leading potentially to 50 expert panel studies in reading and in mathematics, and that NAEP should prepare to carry out these studies. Moreover, high correlations between NAEP scores and state assessment scores are not likely to be seen as satisfactory responses by those who challenge NAEP on alignment grounds.

Although there was near unanimity on the panel concerning the need for studies of alignment with state assessments, there was less agreement on the priority of broader construct validity questions about what NAEP is actually testing, what children do when responding to items, and how NAEP relates to the curriculum taught in schools. There was some concern that NAEP has already set content and performance standards that are based on careful consensus-building processes and that although these are interesting research questions, they do not respond directly to validity challenges. Further, although there was agreement that the proposed studies were aimed at determining specifically what skills are being tested in NAEP, there was some concern that any studies of the relations between what NAEP tests and what is taught in the classroom might take on the appearance of endorsing a standard national curriculum.

Validity of How Skills Are Measured

The NVS panel generally felt that NAEP needs to be able to respond to challenges that contend that the items are not testing the skills described in the framework. NAEP requires more writing in the reading assessment and more reading in the mathematics assessment than do many tests currently in use. This may mean, for example, that students who can read very well but have difficulty writing will not perform as well on the reading assessment as students who are less proficient readers but more facile writers. If this is a substantial effect, then NAEP might tend to corroborate gains of reading programs that simultaneously build writing skills more than of programs that focus on other aspects of reading.[3]

The list of ways that test items can be "contaminated" by unintended skill requirements is long, and the studies needed will have many research questions. Overall, the panel rated the need for these studies as "High," and some panel members indicated that they are "Essential." However, the panel was more convinced of the need for studies of a problem that is currently threatening NAEP: the need for valid assessment of the achievement of students who may be proficient but are burdened by barriers that prevent them from demonstrating their expertise. The question is how to "accommodate" test administration to their special needs without invalidating the results.

A wide variety of accommodations are in use in state assessments, and these are being introduced gradually into NAEP. Some of the more common accommodations are increased time, small-group administration, alternative language forms, and the use of a "scribe" to record responses for students. These accommodations, or non-standard test administrations, make the testing situation feasible for students with disabilities or limited English proficiency by removing specific barriers to the demonstration of their achievement, but they may also make the test easier by removing parts of the target skill domain—skills that students in non-accommodated conditions must also master. For example, if on a mathematics assessment some non-accommodated students with poor writing skills "lose points" that they would have gained if they had been assigned a scribe, then some form of alignment is required. If writing is considered part of mathematics proficiency, then the effect of the scribe must be appropriately subtracted from the score; alternatively, if writing is considered a contaminating skill, then points lost for poor writing skills must somehow be added back in.

---

[3] Although the example of writing in open-ended reading items is used for simplicity, there are other major differences between NAEP and many other assessments currently in use, including NAEP's de-emphasis of vocabulary.

Although research to address this validity question is complex, extensive, and expensive, most members of the NVS panel considered it "Essential" that NAEP carry out studies to determine how to score accommodated performance so that groups of accommodated and non-accommodated students with the same level of content proficiency obtain the same distribution of NAEP scores.

Validity of Who Is Tested

If NAEP is to corroborate state assessment reports of gains at the aggregate level, then NAEP is open to challenge if its sample figures cannot be weighted appropriately to represent the same population of students that the state's own program assesses. This may happen because students in schools not eligible for NAEP participation are included in state testing programs or because selected students fail to participate in NAEP. To comply with P.L. 107-110, states may decide to mandate that schools sampled for NAEP participate, thus solving a past problem of school refusals for NAEP, but home schooling, charter schools, and schools for students with severe disabilities represent areas of concern in representing the full student population. Moreover, some panel members suggested that increasing the stakes for NAEP performance at the school level may result in decreases in participation of the students who would have the most difficulty demonstrating proficiency.

In practice, these population discrepancies may be small, and they may have an even smaller impact on aggregate results if under-represented students tend to perform at approximately the same levels as other students. Although panel members saw the need for each of four component studies (school and student frame completeness and school and student nonparticipation) as only "Moderate," they immediately reacted to this circumstance by indicating that the need for validating the student population *overall* might be more important than any single component. A subsequent rating of the overall problem area indicated a "High" need for these studies. Thus, NAEP should monitor each source of potential population bias and make preparations so that studies to correct for sampling and participation problems can be implemented quickly when signals are received that they are necessary.

The problem of the exclusion of substantial portions of students with disabilities or limited English proficiency from the population represented by NAEP received special attention by the NVS panel. These exclusions have already led to challenges of the validity of NAEP's reports of gains in some states, and the panel members considered the need for research into methods to include "excluded" students in population estimates to be somewhere between "Moderate" and "Essential," most often "High."

Finally, although not specifically included in the rating exercise, NVS panel members suggested related problems during the discussion. These included identifying subgroup sample sizes needed for valid subgroup comparisons, addressing the time-of-year issue with respect to students who do not start the school year near September 1, and, for national NAEP, addressing twelfth-grade participation rates.

Validity of How Data Are Analyzed

The studies proposed by the NVS panel in this area focus on the need to validate analyses that are faster alternatives to past NAEP analyses. The deadline for reporting primary NAEP results will be substantially shorter than NAEP has been able to meet in the past, and shortcuts in analysis may be the only way to meet these deadlines. However, the shortcuts must be shown to produce results that are both valid and consistent with the more extensive analyses that NAEP has employed in the past. The major innovation considered is the production of direct estimates not involving imputed plausible values and not relying on a broad range of contextual information (i.e., "conditioning") to increase the precision of population estimates.

The NVS panel rated the need for such testing of analytical shortcuts between "Moderate" and "Essential," most often "High." Owing to limited time and the recognition of the analytical complexities involved, the panel did not fully discuss or rate two of the studies in this category: 1) omitting the step of subscale estimation from NAEP's primary analyses to provide more stable estimation of a single overall mathematics or reading scale and 2) focusing on estimates for schools, which are by nature more stable than estimates for individual students and which may be sufficient for the purpose of NAEP envisioned in P.L. 107-110. If we "impute" the panel members' ratings on the basis of the assumption that they would reflect the same judgments of need to validate analytical shortcuts as recorded for "direct estimation," then the need to validate these other analytical approaches would also be "High." In fact, these analytical studies are all interrelated and might well be carried out most efficiently as a single overall project.

Finally, the NVS panel considered that NAEP must not be found to be underestimating measurement error and that the component of error due to sampling items from a domain must be estimated. In fact, the development of a standard method for including this error component in overall standard error estimation is under way, and the panel considered that to be prepared to respond to challenges about NAEP's statements of precision and statistical significance, the need for this work is at least "High" and probably "Essential."

Validity of How Results Are Used

The NVS panel's earlier considerations of issues related to consequential validity of NAEP have been overtaken by P.L. 107-110, which calls for a much more concrete and profound use of NAEP results than NAEP has experienced. The NVS panel indicated unanimously that studies are "Essential" to evaluate the validity aspects of whatever operationalization of NAEP's new role is put in place. The panel was careful to distinguish between definition and validation. While recognizing that the need for *defining* the concept of corroboration of state assessment results is a very high priority, the panel acknowledged that this is a policy activity, not a research activity. With regard to *validation*, however, the NVS panel urged that all components of the NAEP design be considered in determining the limits on the validity of the inferences that policymakers might wish to make on the basis of NAEP results.

To emphasize the importance of this area of validity research, the panel returned to this question at the end of the rating session, mentioning studies to ensure that NAEP produces the necessary data to support the inferences that will be made in evaluating state assessment results and rating those studies "Essential."

Perhaps in contrast to preparing valid reports for the Department of Education to consider in reviewing states' results, the panel considered research on ways to maximize information and minimize misinterpretation by members of the public to be only "Moderately" important and studies of ways to help outside analysts avoid erroneous inferences to be of "Low" importance. The latter raised concerns that studies might result in limiting research access to the data.

The most important uses imagined for NAEP in the near future are relative to state assessments, and NAEP should be prepared to work with all the state assessment programs to build a clear understanding of the similarities and differences between what the state is measuring and what NAEP is measuring.

Validity of Trend Reports

Each year the context of NAEP changes: new challenges are presented and new ideas emerge for enhancing NAEP's sampling, measurement, analysis, and reporting. This presents a major problem for NAEP because its primary objective is to measure trends over time, yet we cannot measure change if the yardstick changes at the same time. Nevertheless, change is necessary, and although attempts to measure trends when the yardstick changes must be approximations, the approximations can be dramatically improved by conducting bridge studies to measure the effects of the yardstick change. Typically, these studies present the same participants with alternative (old and new) forms or compare results of old and new analyses of the same data.

An important concern for NAEP is keeping pace with changing concepts of educational achievement. Periodically, therefore, NAEP reconsiders the framework in each content area. In the area of mathematics, for example, two current concerns are how to adapt to the growing availability and use of calculators and computers and how to incorporate higher level mathematics problems in the twelfth-grade assessment. Yet, as these changes are made, there is a need for a bridge study to enable a comparison of the overall level of mathematics proficiency prior to and following changes in the assessment framework, for example, by comparing the performance of the same students on blocks of items taken from earlier assessments and on blocks of new items.

Another concern is the change, now being implemented, from test administration by local school staff to administration by government contractor's staff. Past evaluations of the validity of the NAEP Trial State Assessments found small but statistically significant differences related to test administration factors. Change in the context of NAEP, from low stakes to high stakes for schools, also has an unknown effect, which can distort trends. And upcoming changes in the wording of the questions about race and ethnicity will have unknown effects on the measurement of gaps. Changes in the size of the excluded student population have already been shown to distort trends, and changes in accommodations will have other, unknown effects. Although it is tempting to respond to these changes by saying that NAEP simply cannot measure these trends, the NVS panel rejected this extreme position and deemed bridge studies to be "Essential" whenever there are indications that a change will have a noticeable impact on score distributions.

Thus, NAEP should put in place a system for assessing the potential impact of proposed changes that will guide decisions on the development of corresponding bridge studies.

Another aspect of trend validity is the avoidance of spurious trends between two points in time; the panel considered the value of focusing trend reporting on multiple time points. In this way, more data can be brought to bear on questions about trends over time. The panel considered this to be an area of validity research that is of "High" importance.

Summary

The NAEP Validity Studies panel identified key studies that are needed to address present and imminent threats to the validity and utility of NAEP reports. These studies, described in chapters 2 through 7, cover all phases of assessment, ranging from construct definition and measurement to sampling, analysis, and reporting. The studies described in the preceding chapters overlap in some cases and focus on critical subsets of the issues in other cases. For example, the papers on "what is measured" and "how is it measured" both call for studies to identify deviations between a) what NAEP describes as the domain being tested and b) what is actually tested. And although the paper on special issues related to the measurement of trends focuses on the advantages of using more than two time points in trend analysis, important trend validity issues are also raised by every change in NAEP (e.g., introduction of accommodations, change in race categories, change in stakes) and need to be addressed. In this chapter, the studies have been abstracted and presented together to facilitate priority setting. However, this synthesis does not substitute for a careful reading of the separate chapters on aspects of the assessment.

All the recommended studies would contribute to the state of the art with respect to achievement assessment, fulfilling NAEP's role as a leader in the development of assessment practice. Thus a case can be made for proceeding with each study. The urgency of the studies is not uniform, however: some studies are needed to address threats that have already been made evident, whereas others respond to potential threats that may or may not actually occur.

The aggregate priorities of the validity studies recommended by the NVS panel are shown in Table 8.3. The averages shown in this table are based on the data in Table 8.2, along with the assumption of an equal interval scale. Four studies stand out as essential and two others received ratings that placed them between essential and highly important. Among the remainder, seven studies are evaluated as highly important, and three are of lesser importance or problematic in some way.

Table 8.3–
Aggregate NAEP Validity Research Priority Judgments

Validation Aspect	Study	Average Priority Rating
	Essential	
Uses	17. Meaning of "confirming state results"	4.00
Uses	18. Limits on NAEP's capacity to evaluate state results	4.00
Trends	21. Bridge studies	4.00
Construct	1. Alignment with state standards	3.79
	High to Essential	
Analysis	16. Estimation of item domain sampling error	3.57
Measurement	6. Accommodations	3.54
	High	
Measurement	5. Contaminations	3.25
Sampling	12. Representing excluded SD and LEP students	3.14
Construct	2. Definition: What is being measured	3.07
Sampling	11. Combined studies of population bias	3.00
Analysis	13. Direct estimation with minimal conditioning	3.00
Trends	22. Estimation of multi-time-point trends	3.00
Construct	3. Definition: What students do on the test	2.86
	Not High	
Uses	19. Evaluation of audience interpretations	2.54
Construct	4. Definition: Comparison with curriculum	2.43
Uses	20. Controls and supports for secondary analysis	1.08

Note: Averages are based on scaling responses in Table 8.2 from 0 (Not Needed) and 1 (Low) to 3 (High) and 4 (Essential).

The four studies considered unquestionably **essential** are studies to document the validity of the process of evaluating state assessment results, including a comparison of the alignment of NAEP with state assessment standards and instruments, plus bridge studies to maintain trend reports in the context of changes in measurement. In addition, research to develop valid scoring of accommodated test performance and research to develop methods for adding domain item sampling error to the estimation of measurement error were considered essential by more than half the panel members.

Seven problem areas were considered sufficiently likely to raise serious threats to NAEP's validity to warrant a **high** need for studies to address them. Three of these involve gaining a more thorough understanding of the constructs being assessed by NAEP and the processes that students go through in responding to items, especially identifying irrelevant skill requirements that may contaminate test scores. Two involve monitoring sampling and the methods of representing the student population and developing methods to represent excluded students. The other two highly needed areas involve validating changes in analytic methods that can shorten the time required for analysis and assessing the gains in trend analysis that can be obtained by using data from more than two points in time.

Of the remaining three studies, one, evaluation of (public) audience interpretations of reports, is considered to be **moderately** important, but the other two may be problematic. A comparison of NAEP content with what is taught in the classroom must be carefully constrained so that it does not promote a national curriculum standard, to the detriment of individual state curriculum standards. And developing methods to support certain secondary analyses while suppressing others as misleading, although potentially raising the quality of statistical policy analyses, can also raise issues about freedom of access to government-produced information.

Thus, the recommended NAEP validity research agenda consists of research in four essential areas, nine highly needed areas, and three less important areas. In this phase of setting a validity research agenda, the NVS panel did not consider either the cost of the studies or specific design issues. Some of the studies requiring new data are likely to be the most expensive and to require the greatest time; studies based on expert panel judgments or cognitive lab studies (which gather new information, but not in the same amounts as the "new data" studies) are less expensive and time-consuming, and analyses using existing data in new ways are generally the least expensive and least time-consuming. Nevertheless, there is a wide variation within each category (e.g., how many states will be included in separate alignment studies?), and the costs might range from $20,000 to $200,000 for analytic studies, from $30,000 to $300,000 for expert panel or cognitive lab studies, and from $100,000 to more than $1,000,000 for studies requiring the collection of new data on large samples.

Finally, we note what is *not* included in this set of recommended studies. The NVS panel did not address the issue of conscious cheating, which would certainly invalidate results, primarily because that is an "auditing" rather than a "research" function. And the NVS panel did not focus on issues of developing new frameworks and new topic areas for NAEP, such as testing students' abilities to use computers for writing or their abilities to work as team members. Instead, the panel focused on threats to the validity of NAEP reports in the current context, extended to include the imminent use of NAEP as a check on independent state assessment results.

Appendix to Chapter 4

Exclusions and Accommodations Affect State NAEP Gain Statistics: Mathematics, 1996 to 2000

Don McLaughlin
November 2001

State NAEP reports mean gains for all states that participated in two successive administrations of the same assessment. Those means are based on comparisons of weighted averages of performance of the two samples of students who participated in the two assessments. Standard NAEP reporting has not adjusted gains for changes in percentages of students with disabilities (SD) and students with limited English proficiency (LEP) who are excluded from participation. As a result, artifactual gains have appeared in recent NAEP reports.

The method that NAEP has initiated to address this issue is the inclusion of modified administrations (called "accommodations") for SD and LEP students who otherwise would be excluded from participation. In 2000, there were two NAEP samples in each state, one in which accommodations were not permitted (called the "R2" sample) and one in which accommodations were permitted (called the "R3" sample). At present, NAEP has combined accommodated and non-accommodated scores in the R3 sample without regard to the presence of the accommodation, and as a result, scores of groups of students being accommodated may be artifactually raised.

This report addresses two NAEP validity questions:

1. What would the gains have been if excluded students had been tested, and to what extent are the reported gains biased?
2. To what extent do accommodations raise the scores of SD and LEP participants?

Exclusions and Achievement

Initial analyses focused on the R2 sample in 2000, the sample in which accommodations played no role, and compared performance in that sample to performance in 1996, in a sample in which accommodations also played no role. The NAEP reports of gains for this comparison are presented in Appendix B tables B1 and B2[1], for grades 4 and 8, respectively. A key aspect of these tables is the widespread decrease in percentages of students participating in NAEP, from 1996 to 2000. For example, in Texas, grade 4 participation dropped by 5.1 percent from 1996 to 2000; that is, 5.1 percent more of students selected for NAEP were excluded because they were SD or LEP and judged unable to participate meaningfully in NAEP.

---

[1] All odd numbered tables in this report present grade 4 results. All even numbered tables present corresponding grade 8 results.

Because of this increase in exclusions, one cannot interpret reported NAEP gains as representing gains in achievement because the two assessments, 1996 and 2000, represent different subsets of the student population. To obtain estimates from which one can make gain comparisons, one must estimate the achievement of the excluded students, to add them into full-population estimates. The estimates need not be as accurate as the estimates for students who participate in the NAEP testing sessions, because they constitute a small percentage of the population, but they must not be so imprecise that they raise the standard errors of the population estimates to levels that have an impact on statistical significance testing.

Using a method described elsewhere (McLaughlin, 2000),[2] I developed full-population estimates for the 1996 sample and the 2000 R2 sample. The estimation of full-population achievement means is based on the assumption that excluded students would perform at the same level as included LEP or SD students in the same state with the same demographic characteristics and the same responses to a NAEP SD/LEP questionnaire, which is completed for every SD/LEP student sampled for participation in NAEP. Specifically, a linear regression is carried out, predicting the scores (mean composite plausible values) of included SD and LEP participants from information that is also known about excluded students, and the estimated regression weights are used to estimate the performance of excluded SD and LEP students.

The regression coefficients are shown in tables A1 and A2, for grades 4 and 8. The regression coefficients for the "R2" sample are in the leftmost column of these tables. Fourteen of the predictors are taken from the SD/LEP questionnaire: "Lrn Disab" and the final thirteen. Labels for these variables are included in Appendix B. Those ending in "D" are relevant for students with disabilities, and those labeled "L" are relevant for students with limited English proficiency. The values of these variables are also included in Appendix B, along with recodings used to replace missing data. In each case, a missing response was recoded to indicate that the individual was either not disabled or not limited in English proficiency.

As an example of the interpretation of the entries in table A1, the value –6.883 for RdgGradeD indicates that a teacher's judgment that the student was reading at one grade lower level is associated with a decrement of approximately 7 points on the NAEP scale. The decision to include a variable in the prediction equation was based on regression output listings indicating a value of Student's *t* for the corresponding regression coefficient greater than 1.96.[3] All of the coefficients are in the expected direction, with the exception of SciGradeL, whose sign is probably affected by collinearity with RdgGradeL and MthGradeL.

---

[2] McLaughlin, D.H. (2000). Protecting State NAEP Trends from Changes in SD/LEP Inclusion Rates. Paper presented at the National Institute of Statistical Sciences workshop on NAEP inclusion strategies. Research Triangle Park, NC.

[3] A more sophisticated test of statistical significance is not warranted, because the interpretation of the regression coefficients in tables 1 and 2 is not the aim of this study. Generally, adding the omitted variables would not increase the adjusted R^2, but omitting an included variable would decrease R^2.

The R^2 for predicting the performance of (included) SD and LEP students in the R2 sample was .33. This was compared with the R^2 for the R3 sample, both including and not including the scores of students receiving nonstandard (accommodated) administrations of NAEP, first using the same regression equation used for the prediction in the R2 sample, then using the "best" regression equation.

Table A1. Imputation Regression Coefficients, Grade 4 Math 2000, by Sample.

	R2	R3: including Accommodated		R3: non-Accommodated only	
		Predictors: Same as R2	Best predictors	Predictors: Same as R2	Best predictors
n	*3,939*	*6,639*	*6,639*	*3,515*	*3,515*
R^2	*0.333*	*0.222*	*0.237*	*0.263*	*0.275*
LEP, not SD	8.355	6.235	5.499	8.798	7.539
Minority	–9.532	–9.623	–9.382	–9.558	–9.123
Female	–3.187	–2.628	–2.776	–3.980	–4.043
TITLE 1	–10.087	–3.781	–3.746	–5.292	–5.036
Lrn. Disab.	–4.231	–1.153		–5.182	–4.686
SLUNCH	–5.747	–1.768	–1.771	–1.573	–1.668
PCTBHI	–0.052	–0.154	–0.159	–0.161	–0.179
PCTASN	0.257	0.084	0.118	0.013	
Rdg Grade D	–6.883	–4.259	–3.724	–6.563	–5.941
Mth Grade D	–10.122	–11.383	–7.721	–10.111	–7.499
Mth Curr D			–11.264		–8.665
Sci Grade D			–2.530		
Rdg Part D			–5.176		–10.939
Sci Part D	–8.529	2.219	6.981	–1.565	8.346
Rdg YrsEng			–2.737		–7.096
Mth YrsEng					4.947
Rdg Grade L	–7.310	–6.677	–7.039	–8.173	–8.520
Mth Grade L	–12.316	–0.622		1.532	
Sci Grade L	10.800	–3.142	–3.127	–7.135	–5.199
Rdg Curr L			–5.017		–4.829
Sci Curr L	–8.122	–1.816		–3.150	

In grade 4, the prediction was not as strong in R3 as in R2, especially when accommodated scores were included. However, focusing only on the non-accommodated SD and LEP students in the R3 sample improved the predictability of scores (0.275 vs. 0.237). In grade 8, the prediction in the R3 sample was nearly as good as in the R2 sample, possibly better when the accommodated scores were removed (0.341 vs. 0.324).

The finding that the prediction is weaker when accommodated scores are included is to be expected: accommodations are designed to compensate for the effects of disabilities and limitations. The finding, at grade 8, that the prediction for non-accommodated students in R2 and R3 is roughly equally reliable, can be taken as corroboration of the prediction methodology. On the other hand, the finding at grade 4, that the prediction is weaker for the non-accommodated SD and LEP students in the R3 sample than for similar (non-accommodated) students in the R2 sample (0.333 vs. 0.275), is more difficult to explain. A larger percentage of the R3 sample was accommodated in grade 4 than in grade 8 (47 percent vs. 35 percent), and it may be that the omission of these cases had the effect of reducing the strength of association.

Table A2. Imputation Regression Coefficients, Grade 8 Math 2000, by Sample.

	R2	R3: including Accommodated		R3: non-Accommodated only	
		Predictors: Same as R2	Best predictors	Predictors: Same as R2	Best predictors
n	*3,358*	*5,391*	*5,391*	*3,527*	*3,527*
R^2	*0.324*	*0.298*	*0.304*	*0.330*	*0.341*
IEP	9.561	11.377	11.315	10.613	8.820
DMIN	–9.739	–14.499	–14.913	–15.958	–16.442
DFEM	–5.077	–4.114	–4.174	–4.035	–4.078
TITLE1	–4.671	–9.092	–8.687	–9.159	–8.399
LRNDISAB					–5.075
SLUNCH	–5.931	–4.038	–3.923	–3.262	–3.029
PCTBHI	–0.222	–0.147	–0.144	–0.204	–0.218
PCTASN	0.511	0.421	0.418	0.060	
X012401	–2.616	–4.229	–4.828	–6.323	–6.087
X012501	–16.232	–2.239		–0.569	
X012601	–12.822	–11.621	–9.536	–12.177	–11.192
X012701			–5.320		
X012801					
X012901	–8.812	–2.444		–5.960	–7.013
X013501					
X013701	–8.734	–5.632	–5.789	–8.012	–7.176
X013901					
X014001					
X014201	2.960	–0.356		–0.431	
X014301	–8.610	–9.066	–6.703	–8.775	–5.755
X014401			–8.459		–5.312
X014501	–9.702	0.819	4.777	1.791	
X014801			–6.651		–19.050
X015001	–5.690	–5.725		–8.583	9.393

The purpose for estimating these regression equations is to impute plausible values for the mathematics proficiency of excluded SD and LEP students. The regressions are based on pooled within-state variation, and as a result, the intercept is identically zero in this regression. To impute performance, the intercept is set, for each state, to equate the predicted and actual performance of the included NAEP SD and LEP participants in that state. Thus, the excluded SD and LEP students in each state are predicted to perform better or worse than the included SD and LEP participants in that state to the extent that these groups differ on the variables used in the linear regression equation.

Using this method, I estimate that excluded SD/LEP students perform at a lower level than included SD/LEP students, on average by about 15 points. While this difference varies from-state-to-state, the estimates are sufficiently precise that statistical significance tests for full-population gains can be performed. The results of analyses are shown in tables A3, A4, A5, A6, A7, and A8. Tables A3 and A4 display results for the R2 sample; tables A5 and A6 display results for the subset of non-accommodated students in R3 (accommodated students' scores are set to "missing"); and tables A7 and A8 display results from analyses ignoring information about accommodations (accommodated and non-accommodated scores in the R3 sample are treated the same). Gains significant at the .05 level are indicated by asterisks.

These six tables parallel the four appendix tables B1, B2, B3, and B4, which show figures based on standard NAEP computations.[4] Six comparisons at each grade are of particular interest:

(1) Using R2, how do the standard NAEP results differ from full-population estimates (tables A3 and A4 vs. tables B1 and B2);

(2) Using R3, how do standard NAEP results differ from full-population estimates using non-accommodated SD/LEP students (tables A5 and A6 vs. tables B3 and B4);

(3) How similar are full-population estimates based on two independent samples of non-accommodated SD/LEP students (tables A3 and A4 vs. tables A5 and A6);

(4) How similar are the standard estimates based on two independent samples, one including accommodations (tables B1 and B2 vs. tables B3 and B4);

(5) Using R3, how does including accommodated scores affect full-population estimates (tables A5 and A6 vs. tables A7 and A8); and

(6) How do two different approaches compare (tables A3 and A4 vs. tables B3 and B4)?

---

[4] It should be noted that NAEP has not published gains, as shown in tables A3 and A4, from a non-accommodated administration (e.g., 1996) to an accommodated administration (e.g., 2000/R3).

Full-Population Estimates vs. Standard Estimates: R2

As can be seen from a comparison of the full-population estimates (tables A3 and A4) with standard NAEP estimates (tables B1 and B2), the standard NAEP estimates for gains from 1996 to 2000 are generally greater then the full-population gain estimates, because some of the standard gains are due merely to the increases in percentage of the population excluded.

Based on R2, at grade 4, in 37 states participating in both years, the average full-population gain was 2.3 points on the NAEP scale, but this was inflated to an apparent 3.2 point gain by the increase in the size of the excluded population from 7.4 percent tc 8.9 percent. In six states, the report of whether a significant change occurred would be changed: according to full-population estimates, as opposed to standard NAEP estimates, California experienced a significant gain, whereas Iowa, Missouri, New York, Rhode Island, and Texas did not.

At grade 8, in 35 states participating in both years, the average full-population gain was 1.7 points on the NAEP scale, but this was inflated to an apparent 3.0 point gain by the increase in the size of the excluded population from 6.1 percent to 8.0 percent. In nine states, according to full-population estimates, as opposed to standard NAEP estimates, the following states did not experience a significant gain: Kentucky, Massachusetts, Maryland, Mississippi, Montana, New York, Rhode Island, Vermont, and West Virginia.

Generally, for states in which there was an increase in exclusions from 1996 to 2000, the standard NAEP gains over-estimated what gains would have been if the same population had been represented in both years.

Table A3. Full-Population Estimates of Average Gains from 1996 to 2000, Grade 4, by State, Based on R2.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain (Pct)
Alaska			222.0	1.30			
Alabama	215.0	1.50	209.2	1.31	5.8	2.90 *	0.5
Arkansas	213.7	1.25	212.3	1.38	1.4	0.74	0.1
American Samoa	145.8	5.68					
Arizona	212.7	1.54	212.2	1.85	0.4	0.19	0.6
California	209.8	1.84	202.9	1.78	6.9	2.69 *	6.8
Colorado			221.6	1.04			
Connecticut	230.0	1.17	229.1	1.01	0.9	0.61	–1.9
District of Columbia	189.5	1.25	184.0	0.99	5.5	3.42 *	1.9
DD	225.4	1.22					
Delaware			211.3	0.77			
DO	225.6	0.74					
Florida			211.4	1.24			
Georgia	216.7	1.08	212.7	1.41	4.0	2.27 *	0.7
Guam	179.4	2.03	183.4	1.32	–4.0	–1.66	0.9
Hawaii	210.7	1.19	212.4	1.45	–1.7	–0.88	–4.4
Iowa	229.2	1.32	226.6	1.17	2.6	1.47	–4.5
Idaho	223.8	1.37					
Illinois	221.8	1.97					
Indiana	231.7	1.32	227.3	0.98	4.4	2.65 *	–1.5
Kansas	229.2	1.46					
Kentucky	217.7	1.23	217.2	0.95	0.6	0.35	–2.5
Louisiana	215.1	1.45	206.7	1.13	8.4	4.57 *	0.0
Massachusetts	230.7	1.21	225.3	1.34	5.5	3.04 *	-1.4
Maryland	218.3	1.12	218.2	1.53	0.0	0.01	-1.2
Maine	226.0	0.90	229.1	1.02	–3.0	–2.24 *	-2.6
Michigan	226.1	1.51	223.8	1.25	2.3	1.19	–2.2
Minnesota	232.7	1.43	229.8	1.09	2.9	1.62	0.3
Missouri	224.9	1.18	223.1	1.05	1.8	1.17	–4.8
Mississippi	208.7	1.11	205.5	1.25	3.3	1.95	1.7
Montana	227.7	1.84	225.4	1.25	2.3	1.04	–0.5
North Carolina	230.0	1.04	221.7	1.19	8.3	5.25 *	–6.4
North Dakota	228.6	0.98	229.5	1.24	–0.9	–0.58	–2.2
Nebraska	222.2	2.20	225.1	1.13	-2.9	–1.16	–2.6
New Jersey			225.1	1.37			
New Mexico	208.8	1.34	208.9	1.83	0.0	0.00	–0.4
Nevada	215.3	1.41	213.7	1.34	1.6	0.84	–1.5
New York	220.9	1.53	220.3	1.16	0.6	0.29	-3.7
Ohio	226.3	1.51					
Oklahoma	219.7	1.31					
Oregon	223.5	1.60	219.7	1.36	3.8	1.83	1.0
Pennsylvania			224.6	1.27			
Rhode Island	218.9	1.29	217.7	1.44	1.2	0.60	-5.5
South Carolina	217.5	1.19	210.7	1.14	6.8	4.11 *	–1.8
Tennessee	217.9	1.56	216.5	1.40	1.3	0.64	2.6
Texas	225.1	1.53	224.8	1.31	0.3	0.15	-5.1
Utah	223.2	1.24	223.7	1.09	–0.4	–0.24	–0.9
Virginia	227.2	1.05	221.9	1.36	7.2	4.18 *	–4.1
Virgin Islands	182.0	2.82					
Vermont	226.4	1.50	222.0	1.23	4.5	2.32 *	–4.6
Washington			222.6	1.18			
Wisconsin			228.4	1.08			
West Virginia	219.7	1.22	220.0	1.04	–0.3	–0.21	–1.7
Wyoming	226.7	1.45	221.6	1.33	5.1	2.61 *	–1.9

Table A4. Full-Population Estimates of Average Gains from 1996 to 2000, Grade 8, by State, Based on R2.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain
Alaska							
Alabama	258.8	1.82	252.5	2.18	6.2	2.20 *	2.2
Arkansas	255.4	1.39	256.3	1.60	–0.9	–0.41	–1.2
Arizona	264.9	1.88	263.4	1.54	1.6	0.64	–0.6
American Samoa							
California	256.6	2.69	257.0	1.65	–0.4	–0.13	1.3
Colorado							
Connecticut	276.5	1.67	276.1	1.13	0.4	0.19	–1.9
District of Columbia	229.6	2.13	227.6	1.24	2.0	0.81	0.6
DD							
DO							
Delaware							
Florida							
Georgia	262.6	1.30	258.9	1.61	3.7	1.79	–0.2
Guam	231.3	1.86	237.4	1.69	–6.0	–2.39	–1.4
Hawaii	257.5	1.57	259.1	1.10	–1.6	–0.81	–2.1
Iowa							
Idaho							
Illinois							
Indiana	280.5	1.42	272.4	1.47	8.2	3.99 *	–1.7
Kansas							
Kentucky	265.6	1.23	263.7	1.12	1.8	1.10	–4.8
Louisiana	256.5	1.62	249.7	1.63	6.8	2.94 *	0.2
Massachusetts	277.5	1.55	272.9	2.08	4.6	1.76	–4.1
Maryland	271.2	1.50	266.9	2.13	4.3	1.63	–3.9
Maine	278.8	1.18	281.4	1.31	-2.7	–1.51	–3.8
Michigan	274.1	1.71	273.9	1.82	0.2	0.08	–1.4
Minnesota	284.1	1.44	282.5	1.26	1.6	0.82	–2.5
Missouri	269.0	1.51	269.8	1.33	–0.7	–0.37	–1.5
Mississippi	248.0	1.32	246.2	1.24	1.8	0.98	–0.8
Montana	283.4	1.45	281.0	1.44	2.5	1.20	–2.2
North Carolina	274.9	1.18	265.8	1.44	9.1	4.86 *	–9.4
North Dakota	280.1	1.09	281.9	0.87	–1.8	–1.31	–0.5
Nebraska	278.3	1.34	280.0	1.40	–1.8	–0.92	0.8
New Jersey							
New Mexico	252.2	1.69	258.6	1.30	–6.4	–3.01	–3.8
Nevada							
New York	269.9	2.06	266.4	1.44	3.4	1.36	–5.5
Ohio							
Oklahoma							
Oregon	277.6	1.70	274.0	1.56	3.6	1.55	–2.3
Pennsylvania							
Rhode Island	267.5	1.11	265.2	0.89	2.2	1.57	–4.6
South Carolina	261.7	1.28	256.9	1.39	4.8	2.55*	–1.1
Tennessee	260.8	1.74	260.7	1.35	0.1	0.03	-0.2
Texas	269.6	1.52	265.2	1.53	4.4	2.06*	–0.9
Utah	271.7	1.19	273.6	1.02	–1.9	–1.20	0.2
Virgin Islands							
Virginia	271.7	1.46	266.0	1.43	5.6	2.75*	-2.7
Vermont	278.0	1.11	277.0	1.01	1.0	0.70	–5.3
Washington							
West Virginia	262.7	1.21	260.1	1.11	2.6	1.60	-2.6
Wisconsin							
Wyoming	274.2	1.15	273.5	0.94	0.7	0.46	-2.2

Full-Population Estimates vs. Standard Estimates: R3 in 2000, R2 in 1996

As can be seen from a comparison of the full-population estimates (tables A5 and A6) with standard NAEP estimates (tables B3 and B4), the standard NAEP estimates for gains from 1996 to 2000 are generally smaller then the full-population gain estimates, because R3 in 2000 included more SD/LEP students than R2 in 1996 did. For this comparison, R3-based full-population estimates for 2000 treated accommodated scores as missing and imputed them from relations observed for non-accommodated SD/LEP students' scores.

Based on R3, at grade 4, in 37 states participating in both years, the average full-population gain was 3.3 points on the NAEP scale, but this was deflated to an apparent 2.3 point gain by the decrease in the size of the excluded population from 7.4 percent to 3.7 percent. In six states, there were significant gains according to full-population estimates, as opposed to standard NAEP estimates: California, Connecticut, Georgia, Minnesota, Mississippi, and Texas.

At grade 8, in 35 states participating in both years, the average full-population gain was 2.0 points on the NAEP scale, but this was deflated to a 1.2 point gain by the decrease in the size of the excluded population from 6.1 percent to 4.0 percent. In four states, according to full-population estimates, as opposed to standard NAEP estimates, there were significant gains: Kentucky, Oregon, South Carolina, and West Virginia.

Generally, for states in which there was decrease in exclusions from 1996/R2 to 2000/R3 , the standard NAEP gains over-estimated what gains would have been if the same population had been represented in both years.

Full-Population Estimates Based on R2 vs. Full-Population Estimates Based on Non-Accommodated Cases in R3

As can be seen from a comparison of the R2 full-population estimates (tables A3 and A4) with the R3N (R3 without accommodated scores) full-population estimates (tables A5 and A6), estimates for individual states vary (randomly) between samples. However, at grade 4, there is an overall tendency for R3N full-population estimates to be greater than R2 full-population estimates. Full-population estimates based on R2 and on the non-accommodated students in R3 should be the same, except for random error, if the R2 and R3N samples are equivalent. Therefore, the finding that the grade 4 full-population estimates for 2000 are greater when based on R3N than when based on R2 is of concern. The databases used for estimating the imputation functions are of approximately the same size, as shown in table A1, but the performance of non-accommodated grade 4 SD/LEP students in R3 is higher than their counterparts in R2 (201.4 vs. 198.6). Other factors that contribute to this anomaly of two different full-population estimates (which occurs only for grade 4) are discussed later.

Based on R2, at grade 4, in 37 states participating in both years, the average full-population gain was 2.3 points on the NAEP scale, compared to 3.3 when the 2000 scores were based on R3N. In five states, there were significant gains in full-population estimates based on R3N but not when based on R2: Connecticut, Minnesota, Mississippi, Rhode Island, and Texas. These are, with two exceptions, the same states in which R3N full-population gain estimates are significantly higher than standard NAEP estimates based on R3.

At grade 8, in 37 states participating in both years, the average full-population gain was 1.7 points on the NAEP scale, based on R2, or 2.0 points, based on R3N. In one state, Texas, the full-population gain estimate based on R2 was significant, but not the full-population gain estimate based on R3N; and in four states, the gain based on R3N was significant, while the gain based on R2 was not: Kentucky, Mississippi, Oregon, and West Virginia.

Standard NAEP Estimates Based on R2 and R3

Because the administration in 1996 was parallel to the R2 sample (no accommodations), NAEP has only published indicators of statistically significant gains for the R2 sample in 2000. However, one strategy being considered is to replace R2 with R3 in the future, and to evaluate whether trends can be maintained across a bridge from R2 to R3. It is useful to compare the gains from 1996 to 2000, estimated using R2 and R3 data in 2000. These estimates can be compared between tables B1 and B2 (R2) and tables B3 and B4 (R3). As is expected, gains using the R2 sample are larger, because the 2000 R3 means represent a larger percentage of the SD/LEP population than the 1996 (R2) means do. Even though accommodations are included, the estimated performance of accommodated SD/LEP students tends to be similar to other SD/LEP performances – about one standard deviation lower than the performance of non-SD/LEP students.

Based on R2, ignoring the excluded student population at grade 4, in 37 states participating in both years, the average gain was 3.2 points on the NAEP scale. However, using R3 this was deflated to an apparent 2.3 point gain by the increase in percentage of SD/LEP students included in the represented population, from 9.3 percent to 14.7 percent. In six states, there were significant gains according to standard NAEP R2 gain estimates, as opposed to hypothetical NAEP R3 gain estimates: Georgia, Iowa, Michigan, Missouri, New York, and Texas.

At grade 8, in 35 states participating in both years, the average gain based on the R2 sample was 3.0 points on the NAEP scale, but using R3 this was deflated to an apparent 1.2 point gain by the increase in percentage of SD/LEP students included in the represented population, from 8.1 percent to 12.5 percent. In ten states, there were significant gains according to standard NAEP R2 gain estimates, as opposed to hypothetical NAEP R3 gain estimates: Kentucky, Massachusetts, Maryland, Montana, New York, Rhode Island, South Carolina, Texas, Vermont, and West Virginia.

Generally, for states in which there was a decrease in exclusions from 1996/R2 to 2000/R3, due to introduction of accommodations, the bias in gain estimates was less if gains were measured using R3 in 2000. The fact remains that both estimates are biased by the exclusion of a portion of the population, but the bias is smaller when parts of the excluded population are replaced by estimates that are similar to the performance of other, non-accommodated SD/LEP students.

Table A5. Full-Population Estimates of Average Gains from 1996 to 2000, Grade 4, by State, Based on R3 Non-Accommodated Cases.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain (Pct)
Alaska			222.0	1.30			
Alabama	215.0	1.34	209.2	1.31	5.8	3.09*	0.2
Arkansas	214.0	0.96	212.3	1.38	1.6	0.97	-1.4
American Samoa	149.7	2.95					
Arizona	216.6	1.37	212.2	1.85	4.4	1.90	-0.8
California	210.3	1.60	202.9	1.78	7.4	3.07*	1.9
Colorado			221.6	1.04			
Connecticut	232.2	1.22	229.1	1.01	3.1	1.96*	-0.8
District of Columbia	189.3	1.23	184.0	0.99	5.2	3.31*	-1.1
DD	225.4	1.25					
Delaware			211.3	0.77			
DO	224.7	0.89					
Florida			211.4	1.24			
Georgia	218.0	1.07	212.7	1.41	5.3	2.99*	0.4
Guam	182.1	2.16	183.4	1.32	–1.4	–0.55	1.9
Hawaii	213.1	1.16	212.4	1.45	0.8	0.41	-5.5
Iowa	228.3	1.16	226.6	1.17	1.7	1.03	-3.7
Idaho	222.9	1.29					
Illinois	221.4	1.96					
Indiana	230.5	1.25	227.3	0.98	3.2	2.03*	-3.2
Kansas	231.2	1.63					
Kentucky	217.5	1.31	217.2	0.95	0.3	0.19	-2.0
Louisiana	218.3	1.36	206.7	1.13	11.6	6.56*	-6.1
Massachusetts	231.6	1.14	225.3	1.34	6.3	3.58*	-3.8
Maryland	220.4	1.08	218.2	1.53	2.1	1.13	-0.3
Maine	226.1	1.05	229.1	1.02	–2.9	–2.01*	-3.7
Michigan	227.4	1.41	223.8	1.25	3.7	1.94	-1.5
Minnesota	233.3	1.21	229.8	1.09	3.5	2.17*	-3.5
Missouri	225.8	1.20	223.1	1.05	2.8	1.73	-5.3
Mississippi	209.1	1.19	205.5	1.25	3.6	2.07	1.2
Montana	227.0	1.60	225.4	1.25	1.6	0.78	-2.6
North Carolina	227.7	1.20	221.7	1.19	6.0	3.53*	-6.1
North Dakota	229.3	1.12	229.5	1.24	–0.2	–0.11	-1.8
Nebraska	222.6	1.88	225.1	1.13	–2.5	–1.14	-2.7
New Jersey		.	225.1	1.37			
New Mexico	209.9	1.42	208.9	1.83	1.1	0.46	-3.5
Nevada	215.7	1.14	213.7	1.34	2.1	1.18	-2.9
New York	222.7	1.60	220.3	1.16	2.3	1.18	-6.3
Ohio	228.7	1.41					
Oklahoma	221.7	1.17					
Oregon	221.3	1.73	219.7	1.36	1.6	0.75	-1.7
Pennsylvania		.	224.6	1.27			
Rhode Island	222.9	1.24	217.7	1.44	5.2	2.72*	-7.0
South Carolina	217.4	1.34	210.7	1.14	6.7	3.80*	-4.2
Tennessee	218.7	1.46	216.5	1.40	2.1	1.04	2.5
Texas	228.6	1.14	224.8	1.31	3.9	2.22*	-2.6
Utah	225.2	1.24	223.7	1.09	1.6	0.97	-0.7
Virginia	227.7	1.07	221.9	1.36	7.7	4.43*	-3.9
Virgin Islands	178.8	1.68					
Vermont	230.7	1.63	222.0	1.23	8.8	4.32*	-5.3
Washington			222.6	1.18			
Wisconsin			228.4	1.08			
West Virginia	221.3	1.21	220.0	1.04	1.3	0.80	-2.4
Wyoming	228.1	1.17	221.6	1.33	6.5	3.67*	-3.5

Table A6. Full-Population Estimates of Average Gains from 1996 to 2000, Grade 8, by State, Based on R3, Non-Accommodated Cases.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain
Alaska			275.5	1.76			
Alabama	258.6	1.88	252.5	2.18	6.1	2.11	0.2
Arkansas	255.7	1.39	256.3	1.60	–0.5	–0.25	1.1
American Samoa	192.8	4.89					
Arizona	267.2	1.65	263.4	1.54	3.8	1.69	1.0
California	258.1	2.14	257.0	1.65	1.1	0.40	0.5
Colorado			272.8	1.15			
Connecticut	276.6	1.54	276.1	1.13	0.5	0.27	–1.6
District of Col	229.3	1.59	227.6	1.24	1.7	0.85	–2.4
Delaware			260.8	1.03			
Florida			259.6	1.76			
Georgia	262.3	1.30	258.9	1.61	3.3	1.62	–0.4
Guam	230.5	2.12	237.4	1.69	–6.9	–2.54	–5.0
Hawaii	259.5	1.52	259.1	1.10	0.4	0.20	–2.1
Iowa			280.9	1.37			
Idaho	275.8	1.21					
Illinois	272.2	1.62					
Indiana	280.2	1.38	272.4	1.47	7.8	3.89	–0.7
Kansas	282.1	1.59					
Kentucky	267.4	1.42	263.7	1.12	3.6	2.00	–4.2
Louisiana	257.7	1.59	249.7	1.63	8.0	3.49	–2.8
Massachusetts	277.3	1.40	272.9	2.08	4.4	1.74	–3.5
Maryland	270.2	1.62	266.9	2.13	3.3	1.24	0.2
Maine	280.5	1.15	281.4	1.31	–1.0	–0.54	–2.6
Michigan	275.3	1.85	273.9	1.82	1.5	0.57	–1.0
Minnesota	285.6	1.51	282.5	1.26	3.0	1.55	–1.2
Missouri	268.9	1.53	269.8	1.33	–0.8	–0.42	–2.6
Mississippi	250.0	1.40	246.2	1.24	3.8	2.02	0.0
Montana	284.3	1.37	281.0	1.44	3.4	1.69	–2.4
North Carolina	271.9	1.20	265.8	1.44	6.1	3.27	–7.4
North Dakota	280.4	1.22	281.9	0.87	–1.6	–1.04	–0.3
Nebraska	277.0	1.37	280.0	1.40	–3.0	–1.54	–1.5
New Mexico	254.0	1.70	258.6	1.30	–4.6	–2.14	–3.5
Nevada	262.7	0.93					
New York	267.9	2.11	266.4	1.44	1.5	0.59	–3.6
Ohio	279.2	1.67					
Oklahoma	268.0	1.36					
Oregon	278.5	1.54	274.0	1.56	4.5	2.05	–4.7
Rhode Island	267.3	1.17	265.2	0.89	2.1	1.42	–0.6
South Carolina	262.4	1.41	256.9	1.39	5.5	2.80	–0.5
Tennessee	259.7	1.55	260.7	1.35	–1.1	–0.51	1.3
Texas	268.2	1.62	265.2	1.53	3.0	1.33	–1.3
Utah	271.2	0.99	273.6	1.02	–2.3	–1.64	0.7
Virginia	270.6	1.35	266.0	1.43	4.6	2.35	–2.9
Vermont	278.3	1.35	277.0	1.01	1.3	0.76	–3.2
Washington			272.6	1.39			
Wisconsin			279.4	1.58			
West Virginia	265.4	1.15	260.1	1.11	5.3	3.28	–2.0
Wyoming	274.6	1.00	273.5	0.94	1.1	0.78	–2.3

Full-Population Estimates Including Accommodated Student Scores vs. Imputing Scores for Accommodated Students

NAEP estimates plausible values for performance of accommodated SD/LEP students as if they were not accommodated. Common sense suggests that these scores are likely to be somewhat higher than these same students would have attained without the accommodations. Therefore, imputation of these scores based on background information is likely to produce estimates that are somewhat lower than the NAEP plausible values. As a result, full-population estimates of gains from 1996 to 2000, using the R3 sample, are likely to be somewhat smaller if the accommodated students' scores are imputed (as shown in tables A5 and A6) rather than taken from their accommodated performances (a shown in tables A7 and A8).

In fact, at both grades there are small differences in the estimates, 0.6 points for grade 4 and 0.4 points for grade 8. Gains were estimated to be slightly larger when the accommodated scores were included (without adjustments for the accommodations, of course). Statistical significance was affected for four states at grade 4: Arizona, Iowa, Michigan, and Missouri; but at grade 8, only the District of Columbia was affected.

Because the accommodated sample excludes fewer students from the represented population, it would be preferable to use that sample for imputation, but that requires appropriate estimates of the extent to which the accommodations inflate scores by removing essential component skills from the domain being assessed.

A Practical Comparison: R2-based Full-Population Estimates vs. R3-based Standard NAEP Estimates

Reporting standard results for the R3 sample has been offered by NAEP as an alternative to full-population estimation based on R2. Both of these alternatives estimate lower performance than the standard estimates based on R2, because both ignore fewer SD/LEP students than the standard R2 estimates. This comparison is a bit like comparing apples and oranges, because one method ignores some students while the other does not; and both probably overestimate the performance of some SD/LEP students, but for different reasons: in one case by not adjusting for accommodations and in the other case by statistical regression to the mean due to smaller R^2.

The comparison is between tables A3 and A4 (R2-based full population estimates) and tables B3 and B4 (standard R3-based NAEP estimates). In fact, this comparison identifies the smallest number of deviations in statistical significance of any of the gain comparisons. At grade 4, California and Georgia have significant gains according to the R2-based full-population estimates, while Rhode Island has a significant gain according to standard R3-based estimates. At grade 8, South Carolina and Texas are favored by the R2-based full-population estimates, while Mississippi is favored by the standard R3-based estimates. This small number of differences should not be taken as an indication that the estimates are similar—it is a coincidence based on the simultaneous decrease in participation rates in the R2 sample and increase in participation rates in the R3 sample between 1996 and 2000.

Table A7. Full-Population Estimates of Average Gains from 1996 to 2000, Grade 4, by State, Based on R3 Including Accommodated Cases, in 2000.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain (Pct)
Alaska	.	.	222.0	1.30	.	.	
Alabama	215.8	1.29	209.2	1.31	6.6	3.60*	3.1
Arkansas	214.6	1.07	212.3	1.38	2.3	1.31	2.7
American Samoa	151.2	2.43			.	.	
Arizona	217.9	1.29	212.2	1.85	5.6	2.50*	8.1
California	210.8	1.59	202.9	1.78	7.8	3.29*	10.2
Colorado	.	.	221.6	1.04	.	.	
Connecticut	232.3	1.23	229.1	1.01	3.2	1.99*	3.4
District of Columbia	190.2	1.07	184.0	0.99	6.1	4.23*	6.0
DD							
Delaware			211.3	0.77			
DO	.	.			.	.	
Florida	.	.	211.4	1.24	.	.	
Georgia	218.0	1.04	212.7	1.41	5.3	3.01*	4.3
Guam	182.1	1.89	183.4	1.32	–1.4	–0.60	6.0
Hawaii	212.9	1.03	212.4	1.45	0.5	0.30	–2.8
Iowa	230.4	1.18	226.6	1.17	3.8	2.26*	3.3
Idaho	223.5	1.51			.	.	
Illinois	222.1	1.93			.	.	
Indiana	231.9	1.29	227.3	0.98	4.6	2.85*	2.8
Kansas	231.3	1.65			.	.	
Kentucky	218.7	1.36	217.2	0.95	1.5	0.90	3.1
Louisiana	217.4	1.36	206.7	1.13	10.7	6.04*	5.0
Massachusetts	232.6	1.21	225.3	1.34	7.4	4.08*	6.4
Maryland	220.9	1.13	218.2	1.53	2.6	1.38	5.2
Maine	228.0	1.08	229.1	1.02	–1.1	–0.73	3.0
Michigan	228.1	1.56	223.8	1.25	4.3	2.15*	2.9
Minnesota	232.9	1.28	229.8	1.09	3.2	1.89	3.9
Missouri	227.0	1.15	223.1	1.05	4.0	2.55*	2.3
Mississippi	209.3	1.19	205.5	1.25	3.8	2.22*	3.3
Montana	227.7	1.69	225.4	1.25	2.3	1.10	3.0
North Carolina	228.2	1.22	221.7	1.19	6.4	3.77*	1.8
North Dakota	229.3	1.22	229.5	1.24	–0.2	–0.10	2.1
Nebraska	223.7	1.81	225.1	1.13	–1.4	–0.65	1.6
New Jersey	.	.	225.1	1.37	.	.	
New Mexico	211.3	1.38	208.9	1.83	2.4	1.07	6.3
Nevada	216.7	1.10	213.7	1.34	3.0	1.75	2.1
New York	223.5	1.55	220.3	1.16	3.1	1.61	3.2
Ohio	229.0	1.45			.	.	
Oklahoma	221.7	1.08			.	.	
Oregon	222.7	1.85	219.7	1.36	3.0	1.30	6.1
Pennsylvania	.	.	224.6	1.27	.	.	
Rhode Island	223.1	1.13	217.7	1.44	5.4	2.94*	3.2
South Carolina	217.7	1.33	210.7	1.14	7.1	4.04*	0.5
Tennessee	218.7	1.42	216.5	1.40	2.2	1.09	3.8
Texas	229.2	1.04	224.8	1.31	4.4	2.63*	3.5
Utah	225.7	1.28	223.7	1.09	2.0	1.21	3.0
Virginia	228.2	1.02	221.9	1.36	8.2	4.80*	2.7
Virgin Islands	179.4	1.65			.	.	
Vermont	230.4	1.63	222.0	1.23	8.5	4.19*	3.3
Washington	.	.	222.6	1.18	.	.	
Wisconsin	.	.	228.4	1.08	.	.	
West Virginia	222.2	1.22	220.0	1.04	2.2	1.38	5.6
Wyoming	227.8	1.11	221.6	1.33	6.2	3.60*	2.3

Table A8. Full-Population Estimates of Average Gains from 1996 to 2000, Grade 8, by State, Based on R3, Including Accommodated Cases, in 2000.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain
Alaska			275.5	1.76			
Alabama	259.2	1.84	252.5	2.18	6.6	2.32*	0.7
Arkansas	255.8	1.49	256.3	1.60	–0.5	–0.23	4.8
American Samoa	190.9	5.08					
Arizona	267.2	1.62	263.4	1.54	3.8	1.71	5.5
California	258.2	2.09	257.0	1.65	1.1	0.42	5.8
Colorado			272.8	1.15			
Connecticut	277.0	1.42	276.1	1.13	0.9	0.51	2.2
District of Columbia	232.0	1.31	227.6	1.24	4.4	2.46*	3.5
DD	272.3	2.47					
Delaware			260.8	1.03			
DO	277.4	1.03					
Florida			259.6	1.76			
Georgia	262.9	1.33	258.9	1.61	4.0	1.93	2.2
Guam	230.7	2.53	237.4	1.69	–6.7	–2.20*	–3.3
Hawaii	259.8	1.54	259.1	1.10	0.7	0.37	–0.1
Iowa			280.9	1.37			
Idaho	276.0	0.99					
Illinois	272.0	1.52					
Indiana	279.9	1.33	272.4	1.47	7.5	3.81*	2.5
Kansas	281.4	1.65					
Kentucky	267.8	1.34	263.7	1.12	4.0	2.30*	0.2
Louisiana	257.8	1.48	249.7	1.63	8.0	3.65*	3.4
Massachusetts	277.8	1.42	272.9	2.08	4.9	1.95	5.3
Maryland	270.4	1.71	266.9	2.13	3.5	1.29	4.0
Maine	280.0	1.09	281.4	1.31	–1.4	–0.83	2.1
Michigan	275.2	1.90	273.9	1.82	1.4	0.52	1.3
Minnesota	286.1	1.51	282.5	1.26	3.5	1.79	1.3
Missouri	269.6	1.38	269.8	1.33	–0.2	–0.12	4.2
Mississippi	250.8	1.38	246.2	1.24	4.6	2.46*	1.2
Montana	284.0	1.51	281.0	1.44	3.1	1.47	0.9
North Carolina	273.7	1.31	265.8	1.44	7.9	4.05*	–0.6
North Dakota	280.8	1.09	281.9	0.87	–1.2	–0.83	1.7
Nebraska	277.7	1.46	280.0	1.40	–2.4	–1.16	0.8
New Mexico	255.2	1.65	258.6	1.30	–3.4	–1.60	0.5
Nevada	262.6	0.87					
New York	269.1	2.24	266.4	1.44	2.7	1.01	3.6
Ohio	278.9	1.59					
Oklahoma	267.8	1.32					
Oregon	278.8	1.50	274.0	1.56	4.8	2.20*	1.4
Rhode Island	267.2	1.21	265.2	0.89	1.9	1.29	3.8
South Carolina	262.1	1.39	256.9	1.39	5.2	2.67*	1.8
Tennessee	259.9	1.44	260.7	1.35	–0.8	–0.41	1.9
Texas	268.5	1.51	265.2	1.53	3.3	1.54	0.7
Utah	271.8	1.01	273.6	1.02	–1.8	–1.22	3.4
Virginia	271.3	1.27	266.0	1.43	5.2	2.74*	1.0
Vermont	278.8	1.46	277.0	1.01	1.8	1.03	1.2
Washington			272.6	1.39			
Wisconsin			279.4	1.58			
West Virginia	264.7	1.20	260.1	1.11	4.6	2.80*	5.8
Wyoming	274.9	1.02	273.5	0.94	1.4	0.99	0.8

Aggregate Sample Comparisons for 2000

Although there is noticeable random sampling variation in the SD/LEP scores in any single state, it is desirable that, across the aggregate of participating states, the full-population estimates based on different samples be similar. In 2000, there were three samples on which full-population estimates can be based: R2, R3 without the accommodated scores, and R3 (including accommodated scores). The estimates for these three samples are shown in tables A9 and A10. Note that these values are different from values mentioned earlier, because they are based on all participating states, not limited to those that also participated in 1996.

The results for grade 8 (table A10) are as expected: 268.0 or 268.1, and slightly higher, 268.4, when accommodated scores are included. The results for grade 4 (table A9), on the other hand, are not so consistent: the full-population estimate based on R3 (without accommodated scores) is 1.1 points higher than the estimate based on R2. It appears that the 10.9 percent of imputed scores in R3N (R3 excluding accommodated scores) are noticeably higher than the 9.6 percent of imputed scores in R2 (191.5 vs. 182.3). There appear to be three reasons for this: (1) the unimputed SD/LEP scores, which "anchor" the imputed scores, are 2.8 points higher in R3N; (2) the students with accommodated scores in R3 have a higher level of proficiency than similar percentages of students in R2, as indicated by the fact that their mean performance is the same as the non-accommodated students in R3 (201.4); and (3) the R^2 is smaller for R3N (.275 vs. .333), leading to greater regression to the "anchor." These three factors create the "anomaly" that the two full-population estimates differ by 1.1 points.

In this context, the R2 value is preferred because (a) it is based on A slightly larger database and (b) it is based on a slightly higher R^2. Although the addition of accommodations increases the numbers of students who actually participate in NAEP testing sessions, none of the values without imputations, whether based on R2 or R3, are sufficiently close to the R2 full-population estimates to warrant their use as full-population estimates. Moreover, the published R3 estimates (which include accommodated scores) are inflated to an unknown extent by the accommodations, to which we turn next.

Table A9. Average of Grade 4 Achievement Estimates for 2000, by Sample.

	R2		R3 (non-accommodated only)		R3	
	Mean	Percent	Mean	Percent	Mean	Percent
Non-SD/LEP	227.6	(81.1)	227.6	(81.0)	227.6	(81.0)
Included SD/LEP	198.6	(9.3)	201.4	(8.1)	201.4	(14.7)
Imputed SD/LEP	182.3	(9.6)	191.5	(10.9)	189.4	(4.2)
Full-Population Average	220.5	(100.0)	221.6	(100.0)	222.2	(100.0)
Average without Imputations	224.6	(90.4)	225.2	(89.1)	223.6	(95.7)

Table A10. Average of Grade 8 Achievement Estimates for 2000, by Sample.

	R2		R3 (non-accommodated only)		R3	
	Mean	Percent	Mean	Percent	Mean	Percent
Non-SD/LEP	276.3	(83.1)	276.3	(83.1)	276.3	(83.1)
Included SD/LEP	236.7	(8.1)	234.4	(8.4)	232.9	(12.5)
Imputed SD/LEP	218.1	(8.7)	219.9	(8.4)	217.5	(4.4)
Full-Population Average	268.0	(100.0)	268.1	(100.0)	268.4	(100.0)
Average without Imputations	272.8	(91.2)	272.5	(91.5)	270.7	(95.6)

Effects of Accommodations

An essential criterion for valid comparison of test scores of two groups is that they be given either the same test or a parallel (i.e., functionally perfectly equivalent) test. A non-standard administration of a test is *prima facie* a different test, so comparing scores of two groups of students, in which some members of one group had a non-standard test administration, is not valid, **unless** the non-standard administration is shown to be functionally equivalent to the standard administration. An example of an obvious non-standard but equivalent test administration is the provision to accommodate a paraplegic student by allowing a wheelchair to be used instead of a chair and desk, in a test of reading or mathematics. It would not be equivalent, of course, if the test were of running speed.

Carefully controlled, randomized experiments are needed to determine the effect of particular accommodations (i.e., non-standard test administrations) on test scores. Students with and without disabilities, but with the same level of proficiency (determined externally) must be assigned to standard and non-standard test administration conditions to determine how the non-standard test administration should be scored to yield equivalent proficiency estimates.

In the absence of such studies, it is possible to use the NAEP data to estimate the extent to which variations in NAEP scores are associated with accommodations, for SD/LEP students with the same *predicted* performance. For this, we expand the regression analyses used in the preceding sections to include indicators of types of accommodations. The results, shown in tables A11 and A12, indicate the number of NAEP points associated with each of the accommodations, which had apparently significant effects.

The effects are quite different between grades 4 and 8. At grade 4, the LEP accommodations were not effective, but at grade 8, the bilingual glossary was quite helpful. At grade 4, accommodations for students with disabilities were more effective: reading aloud, small group, one-on-one, and "scribed" administrations. The one-on-one administration was also effective at grade 8. And at both grades, miscellaneous types of accommodations were associated with higher scores.

Table A11. Imputation Coefficients for Accommodation Types, Grade 4 Math 2000.

Predictor	Coefficient	Predictor	Coefficient
n	*6,639*		
R^2	*0.251*	*Accommodations*	
LEP, not SD	5.730		
Minority	–9.175	Bilingual Book	–3.979
Female	–2.935	Read Aloud	5.874
TITLE 1	–4.483	Small Group	8.572
SLUNCH	–1.816	One-on-One	12.274
PCTBHI	–0.129	Scribe/PC	34.042
PCTASN	0.108	Other	10.881
RdgGradeD	–4.268		
MthGradeD	–7.875		
MthCurrD	–11.003		
SciGradeD	–3.102		
RdgPartD	–6.406		
SciPartD	5.074		
RdgYrsEng	–2.849		
RdgGradeL	–7.228		
SciGradeL	–3.850		
RdgPartL	–4.403		

Table A12. Imputation Coefficients for Accommodation Types, Grade 8 Math 2000.

Predictor	Coefficient	Predictor	Coefficient
n	*5,391*		
R^2	*0.312*		
LEP, not SD	10.712	*Accommodations*	
Minority	–14.861		
Female	–4.116	Bilingual Dictionary	22.372
TITLE 1	–8.783	Extended Time	–7.591
SLUNCH	–3.566	One-on-One	5.729
PCTBHI	–0.145	Other	8.133
PCTASN	0.280		
RdgGradeD	–4.897		
MthGradeD	–9.552		
MthCurrD	–5.659		
SciPartD	–5.234		
RdgGradeL	–7.521		
MthGradeL	–7.306		
SciGradeL	4.804		
RdgPartL	–7.436		

A more careful analysis, suggested by John Mazzeo, is to use the prediction equation based on SD/LEP students who had the standard NAEP administration (i.e., were not accommodated) and carry out a t-test on the residual performance of students who had each type of accommodation. The results are shown in tables A13 and A14. These results follow the same pattern as described above, although the only significantly effective accommodation at grade 8 is the bilingual glossary.

It is tempting to apply the differences noted in tables A13 and A14 to adjust scores of students who received accommodations to approximate what their scores would have been without the accommodations. However, it's not that simple. To see this, we focus on two of the accommodations in grade 8: the bilingual glossary and extended time.

The scores of the LEP students who had access to bilingual glossaries appear to have been 28.9 points higher than the scores of comparable students who had the standard administration, but it does not make sense to subtract 28.9 points from their scores, even if that value was determined without error. That is because we think that the mathematics skills NAEP is attempting to assess do not include being able to read fluently in English. Providing the glossary does not dilute the mathematics skill requirements for solving problems on the test. (If the test were of English reading achievement, the picture would be different, of course.) Assuming that the bilingual glossary does not, coincidentally, include some mathematical definitions (e.g., what parallel lines are) that substitute for a component of the substance of mathematical knowledge, this accommodation would appear to be both effective and appropriate.

The recommendation for NAEP would be to include the availability of a bilingual glossary as part of the standard administration for all sessions, to ensure that comparisons are not biased because some LEP students have the bilingual glossary and others do not. For comparisons that do not have that standard (e.g., gains from 1996 to 2000), scores of sessions that did not allow accommodations (e.g., R2) should be used.

Turning now to extended time, the scores of grade 8 students who had this non-standard administration were 4.2 points *lower* than the scores of similar students. (Although not statistically significant, it probably would have been in a larger sample.) Although some accommodations might actually hurt performance, it is nearly inconceivable that this accommodation caused lower scores. Much more likely is the possibility that students who were given extended time were not as proficient in mathematics as expected, due to some unmeasured characteristic, not included in the NAEP descriptive questionnaire for SD and LEP students. In this case, one would just ignore the 4.2-point effect (even if it were measured without error), rather than inflate (!) the scores of those who had extended time by that amount.

But then we are left with doubt about all of the coefficients: Were the accommodated and non-accommodated students with the same disability and background profiles really equivalent? It is essential that research be done in which students are randomly assigned to accommodated and standard administrations, to ensure that, at least on average, equivalent groups of students are available for these analyses. Use of existing NAEP data can suggest directions for research, but they cannot answer the question of how to score the performance of accommodated students to construct valid comparisons.

Table A13. Mean Effects of Accommodations, Grade 4.

Accommodation Type	Effect	n	Student's t
Bilingual Book	–2.8	364	–1.49
Bilingual Glossary	6.1	3	—
Large Print	10.7	16	1.88
Extended Time	–0.8	359	–0.31
Read Aloud	7.4	475	3.63
Small Group	9.7	1522	9.97
One-on-One	14.5	278	5.95
Scribe/PC	37.1	27	2.96
Other	12.2	80	2.65

Table A14. Mean Effects of Accommodations, Grade 8.

Accommodation Type	Effect	n	Student's t
Bilingual Book	2.1	116	0.40
Bilingual Glossary	28.9	25	2.83
Large Print	–3.9	6	–0.24
Extended Time	–4.2	515	–1.32
Read Aloud	2.7	223	0.76
Small Group	2.9	805	1.56
One-on-One	8.1	111	1.51
Scribe/PC	–5.1	8	–0.31
Other	9.7	55	1.18

Supplement to the Appendix

Table A15. NAEP Standard Estimates of Average Gains from 1996 to 2000, Grade 4, by State, Based on R2.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain
Alaska			223.8	1.26			
Alabama	217.9	1.41	211.6	1.24	6.3	3.36 *	0.5
Arkansas	217.1	1.13	215.8	1.46	1.2	0.66	0.1
American Samoa	156.7	3.90					
Arizona	218.8	1.42	217.6	1.73	1.2	0.53	0.6
California	213.6	1.84	209.1	1.84	4.4	1.70	6.8
Colorado			225.8	1.04			
Connecticut	234.2	1.16	232.0	1.10	2.2	1.39	–1.9
District of Columbia	193.3	1.17	187.1	1.09	6.2	3.85 *	1.9
DD	228.0	1.18					
DO			215.0	0.64			
Delaware	227.6	0.73					
Florida			215.8	1.16			
Georgia	219.6	1.06	215.5	1.49	4.1	2.24 *	0.7
Guam	184.3	2.34	188.4	1.27	–4.1	–1.53	0.9
Hawaii	215.9	1.07	215.0	1.45	0.9	0.50	–4.4
Iowa	232.9	1.27	229.1	1.08	3.8	2.26 *	–4.5
Idaho	226.9	1.21					
Illinois	224.9	1.92					
Indiana	234.4	1.08	229.4	1.05	5.0	3.34 *	–1.5
Kansas	232.0	1.53					
Kentucky	221.0	1.17	220.0	1.07	1.0	0.63	–2.5
Louisiana	218.0	1.40	209.0	1.11	8.9	5.02 *	0.0
Massachusetts	235.0	1.12	229.0	1.35	6.0	3.41 *	–1.4
Maryland	222.3	1.27	220.7	1.56	1.6	0.80	–1.2
Maine	230.6	0.92	232.2	1.02	–1.6	–1.19	–2.6
Michigan	230.9	1.43	226.3	1.27	4.6	2.42 *	–2.2
Minnesota	235.3	1.32	232.2	1.08	3.1	1.80	0.3
Missouri	228.6	1.19	224.7	1.07	3.8	2.39 *	–4.8
Mississippi	211.0	1.07	208.4	1.22	2.5	1.57	1.7
Montana	229.8	1.81	227.5	1.23	2.3	1.05	–0.5
North Carolina	232.5	1.00	224.3	1.19	8.1	5.24 *	–6.4
North Dakota	230.9	0.86	230.9	1.23	0.0	–0.01	–2.2
Nebraska	225.9	1.72	227.5	1.18	–1.6	–0.77	–2.6
New Jersey			227.2	1.49			
New Mexico	213.9	1.48	213.8	1.75	0.0	0.01	–0.4
Nevada	220.3	1.18	217.6	1.30	2.7	1.51	–1.5
New York	226.6	1.33	222.6	1.24	3.9	2.16 *	–3.7
Ohio	230.6	1.33					
Oklahoma	225.0	1.26					
Oregon	226.6	1.64	223.5	1.35	3.2	1.49	1.0
Pennsylvania			226.2	1.23			
Rhode Island	224.6	1.22	220.4	1.39	4.2	2.27 *	–5.6
South Carolina	220.4	1.39	213.2	1.30	7.2	3.80 *	–1.8
Tennessee	219.8	1.49	219.2	1.40	0.7	0.32	2.6
Texas	232.7	1.21	228.7	1.36	4.0	2.17 *	–5.1
Utah	227.3	1.22	226.5	1.15	0.8	0.46	–0.9
Virginia	230.4	1.27	222.6	1.36	7.8	4.16 *	–4.1
Virgin Islands	182.9	2.81					
Vermont	231.7	1.63	224.9	1.22	6.8	3.35 *	–4.6
Washington			225.1	1.24			
Wisconsin			231.4	0.96			
West Virginia	224.8	1.20	223.4	1.01	1.5	0.96	–1.7
Wyoming	229.3	1.30	223.2	1.38	6.1	3.19 *	–1.9

Table A16. NAEP Standard Estimates of Average Gains from 1996 to 2000, Grade 8, by State, Based on R2.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain
Alaska							
Alabama	262.2	1.77	256.6	2.15	5.6	2.00 *	2.2
Arkansas	261.4	1.37	261.7	1.52	–0.3	–0.14	–1.2
Arizona	270.7	1.53	267.9	1.56	2.8	1.30	–0.6
American Samoa							
California	262.2	2.04	262.8	1.85	–0.6	–0.22	1.3
Colorado							
Connecticut	281.9	1.37	279.6	1.12	2.3	1.31	–1.9
District of Columbia	234.4	2.19	232.8	1.35	1.6	0.62	0.6
DD							
DO							
Delaware							
Florida							
Georgia	266.3	1.25	262.5	1.65	3.9	1.87	–0.2
Guam	233.5	2.15	238.6	1.68	–5.2	–1.89	–1.4
Hawaii	262.3	1.36	262.1	0.97	0.2	0.13	–2.1
Iowa							
Idaho							
Illinois							
Indiana	283.1	1.45	275.5	1.44	7.5	3.68 *	–1.7
Kansas							
Kentucky	271.6	1.40	266.6	1.07	5.0	2.82 *	–4.8
Louisiana	259.0	1.50	252.4	1.57	6.6	3.04 *	0.2
Massachusetts	283.1	1.25	277.6	1.74	5.6	2.59 *	–4.1
Maryland	276.0	1.43	269.7	2.13	6.3	2.47 *	–3.9
Maine	283.6	1.19	284.1	1.29	–0.4	–0.24	–3.8
Michigan	278.5	1.60	276.9	1.79	1.6	0.66	–1.4
Minnesota	287.7	1.44	284.0	1.34	3.6	1.83	–2.5
Missouri	273.6	1.46	273.3	1.39	0.3	0.15	–1.5
Mississippi	254.0	1.30	250.2	1.19	3.8	2.17 *	–0.8
Montana	286.6	1.22	283.0	1.30	3.6	2.00 *	–2.2
North Carolina	280.1	1.13	267.8	1.42	12.3	6.77 *	–9.4
North Dakota	283.1	1.07	284.2	0.91	–1.1	–0.82	–0.5
Nebraska	280.6	1.12	282.8	1.02	–2.1	–1.42	0.8
New Jersey							
New Mexico	259.8	1.74	262.0	1.22	–2.1	–1.00	–3.8
Nevada							
New York	276.3	2.09	270.2	1.66	6.0	2.26 *	–5.5
Ohio							
Oklahoma							
Oregon	280.6	1.65	275.3	1.47	4.3	1.95	–2.3
Pennsylvania							
Rhode Island	273.4	1.11	268.9	0.92	4.6	3.15 *	–4.6
South Carolina	266.4	1.39	260.8	1.54	5.6	2.69 *	–1.1
Tennessee	263.4	1.72	263.1	1.40	0.3	0.14	–0.2
Texas	274.8	1.47	270.2	1.43	4.6	2.26 *	–0.9
Utah	275.4	1.16	276.8	1.03	–1.3	–0.86	0.2
Virgin Islands							
Virginia	276.7	1.50	269.8	1.56	6.9	3.19 *	–2.7
Vermont	283.4	1.10	279.3	0.95	4.2	2.87 *	–5.3
Washington							
West Virginia	270.8	1.00	264.9	1.02	5.9	4.13 *	–2.6
Wisconsin							
Wyoming	276.7	1.18	274.8	0.91	1.9	1.29	–2.2

Table A17. NAEP Standard Estimates of Average Gains from 1996 to 2000, Grade 4, by State, Based on R3 in 2000 and R2 in 1996.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain
Alaska	.	.	223.8	1.26	.	.	
Alabama	217.2	1.18	211.6	1.24	5.6	3.26*	3.1
Arkansas	216.2	1.11	215.8	1.46	0.3	0.18	2.7
American Samoa	151.9	4.04			.	.	
Arizona	218.9	1.27	217.6	1.73	1.3	0.60	8.1
California	212.7	1.63	209.1	1.84	3.6	1.45	10.2
Colorado	.	.	225.8	1.04	.	.	
Connecticut	233.8	1.14	232.0	1.10	1.7	1.10	3.4
District of Columbia	191.6	1.28	187.1	1.09	4.4	2.64*	6.0
DD							
DO			215.0	0.64			
Delaware	.	.			.	.	
Florida	.	.	215.8	1.16	.	.	
Georgia	219.0	1.11	215.5	1.49	3.5	1.89	4.3
Guam	184.5	2.03	188.4	1.27	−3.9	−1.63	6.0
Hawaii	216.3	1.10	215.0	1.45	1.4	0.74	−2.8
Iowa	231.1	1.26	229.1	1.08	1.9	1.17	3.3
Idaho	224.5	1.22			.	.	
Illinois	223.0	1.94			.	.	
Indiana	233.0	1.08	229.4	1.05	3.6	2.41*	2.8
Kansas	232.1	1.62			.	.	
Kentucky	219.4	1.28	220.0	1.07	−0.6	−0.36	3.1
Louisiana	218.2	1.32	209.0	1.11	9.2	5.31*	5.0
Massachusetts	233.4	1.20	229.0	1.35	4.4	2.45*	6.4
Maryland	221.5	1.16	220.7	1.56	0.8	0.43	5.2
Maine	229.5	0.92	232.2	1.02	−2.7	−1.93	3.0
Michigan	229.3	1.58	226.3	1.27	3.1	1.50	2.9
Minnesota	233.7	1.28	232.2	1.08	1.5	0.91	3.9
Missouri	227.8	1.28	224.7	1.07	3.1	1.84	2.3
Mississippi	210.6	1.17	208.4	1.22	2.1	1.26	3.3
Montana	228.5	1.69	227.5	1.23	1.0	0.46	3.0
North Carolina	229.9	1.12	224.3	1.19	5.6	3.40*	1.8
North Dakota	229.8	1.06	230.9	1.23	−1.1	−0.67	2.1
Nebraska	225.1	1.85	227.5	1.18	−2.5	−1.13	1.6
New Jersey	.	.	227.2	1.49	.	.	
New Mexico	213.5	1.42	213.8	1.75	−0.4	−0.16	6.3
Nevada	219.6	1.03	217.6	1.30	1.9	1.16	2.1
New York	225.1	1.41	222.6	1.24	2.5	1.34	3.2
Ohio	230.0	1.50			.	.	
Oklahoma	223.7	1.03			.	.	
Oregon	223.9	1.80	223.5	1.35	0.4	0.19	6.1
Pennsylvania	.	.	226.2	1.23	.	.	
Rhode Island	224.1	1.06	220.4	1.39	3.7	2.09*	3.2
South Carolina	219.9	1.30	213.2	1.30	6.7	3.64*	0.5
Tennessee	219.8	1.43	219.2	1.40	0.7	0.33	3.8
Texas	231.3	1.07	228.7	1.36	2.6	1.48	3.5
Utah	226.8	1.28	226.5	1.15	0.3	0.17	3.0
Virginia	229.5	0.99	222.6	1.36	6.9	4.09*	2.7
Virgin Islands	181.4	2.99			.	.	
Vermont	231.6	1.57	224.9	1.22	6.7	3.37*	3.3
Washington	.	.	225.1	1.24	.	.	
Wisconsin	.	.	231.4	0.96	.	.	
West Virginia	223.2	1.16	223.4	1.01	−0.1	−0.10	5.6
Wyoming	228.6	1.12	223.2	1.38	5.4	3.05*	2.3

Note: The rightmost three columns in this table have not been published by NAEP but are presented here for comparison with other tables in this report.

Table A18. NAEP Standard Estimates of Average Gains from 1996 to 2000, Grade 8, by State, Based on R3 in 2000 and R2 in 1996.

State	2000 Mean	2000 Std.Error	1996 Mean	1996 Std. Error	Gain	t_{Gain}	Inclusion Gain
Alaska			277.6	1.77			
Alabama	263.6	1.83	256.6	2.15	7.0	2.49*	0.7
Arkansas	257.4	1.47	261.7	1.52	–4.2	–2.00*	4.8
American Samoa	192.2	5.46					
Arizona	268.6	1.76	267.9	1.56	0.7	0.30	5.5
California	259.8	2.12	262.8	1.85	–3.0	–1.06	5.8
Colorado			275.6	1.09			
Connecticut	280.8	1.26	279.6	1.12	1.2	0.71	2.2
District of Col	234.6	1.05	232.8	1.35	1.8	1.05	3.5
Delaware			266.7	0.95			
Florida			263.6	1.84			
Georgia	265.4	1.23	262.5	1.65	2.9	1.41	2.2
Guam	233.6	2.59	238.6	1.68	–5.0	–1.62	–3.3
Hawaii	262.2	1.41	262.1	0.97	0.1	0.03	–0.1
Iowa			284.0	1.31			
Idaho	277.2	1.02					
Illinois	274.5	1.67					
Indiana	281.3	1.36	275.5	1.44	5.8	2.92*	2.5
Kansas	283.0	1.69					
Kentucky	269.9	1.29	266.6	1.07	3.3	1.96	0.2
Louisiana	258.6	1.46	252.4	1.57	6.2	2.88*	3.4
Massachusetts	278.9	1.45	277.6	1.74	1.4	0.61	5.3
Maryland	271.9	1.74	269.7	2.13	2.3	0.82	4.0
Maine	281.4	1.10	284.1	1.29	–2.7	–1.60	2.1
Michigan	277.3	1.90	276.9	1.79	0.4	0.16	1.3
Minnesota	287.0	1.39	284.0	1.34	2.9	1.52	1.3
Missouri	270.9	1.46	273.3	1.39	–2.4	–1.17	4.2
Mississippi	254.1	1.08	250.2	1.19	3.9	2.43*	1.2
Montana	285.2	1.37	283.0	1.30	2.2	1.18	0.9
North Carolina	276.2	1.28	267.8	1.42	8.4	4.39*	–0.6
North Dakota	281.9	1.12	284.2	0.91	–2.3	–1.62	1.7
Nebraska	280.0	1.21	282.8	1.02	–2.8	–1.74	0.8
New Mexico	259.3	1.33	262.0	1.22	–2.6	–1.46	0.5
Nevada	264.9	0.84					
New York	271.5	2.19	270.2	1.66	1.2	0.45	3.6
Ohio	280.6	1.57					
Oklahoma	269.7	1.29					
Oregon	280.1	1.52	276.3	1.47	3.7	1.76	1.4
Rhode Island	268.9	1.26	268.9	0.92	0.1	0.04	3.8
South Carolina	264.6	1.50	260.8	1.54	3.8	1.77	1.8
Tennessee	261.6	1.49	263.1	1.40	–1.5	–0.73	1.9
Texas	273.4	1.65	270.2	1.43	3.2	1.49	0.7
Utah	273.5	1.15	276.8	1.03	–3.2	–2.10*	3.4
Virginia	274.8	1.30	269.8	1.56	5.0	2.46*	1.0
Vermont	280.5	1.52	279.3	0.95	1.2	0.70	1.2
Washington			276.1	1.28			
Wisconsin			282.8	1.53			
West Virginia	266.5	1.25	264.9	1.02	1.6	1.00	5.8
Wyoming	275.6	0.98	274.8	0.91	0.8	0.59	0.8

Note: The rightmost three columns in this table have not been published by NAEP but are presented here for comparison with other tables in this report.

Table A19. Variables from the SD/LEP Questionnaire Used in the Linear Regressions

```
X012201  SEVERITY LABEL='DEGREE OF STUDENT''S DISABILITY         '
X012401  RDGGRADD LABEL='ENGLISH GRADE LEVEL STUDENT RECEIVING RE'
X012501  RDGCURR  LABEL='CURRICULUM SAME AS NONDISABLED READING/L'
X012601  MTHGRADD LABEL='ENGLISH GRADE LEVEL STUDENT RECEIVING MA'
X012701  MTHCURR  LABEL='CURRICULUM SAME AS NONDISABLED IN MATHEM'
X012801  SCIGRADD LABEL='ENGLISH GRADE LEVEL STUDENT RECEIVING SC'
X012901  SCICURR  LABEL='CURRICULUM SAME AS NONDISABLED IN SCIENC'
X013501  RDGPARTD LABEL='HOW PARTICIPATE IN NAEP READING LANGUAGE'
X013701  SCIPARTD LABEL='HOW PARTICIPATE IN NAEP SCIENCE         '
X013901  RDGYRSEN LABEL='YEARS RECEIVING ACADEMIC INSTRUCTION IN '
X014001  MTHYRSEN LABEL='YEARS RECEIVING ACADEMIC INSTRUCTION MAT'
X014101  SCIYRSEN LABEL='YEARS RECEIVING ACADEMIC INSTRUCTION IN '
X014201  PCTTIME  LABEL='THIS YEAR PERCENT ACADEMIC INSTRUCTION N'
X014301  RDGGRADL LABEL='GRADE LEVEL RECEIVING ENGLISH READING/LA'
X014401  MTHGRADL LABEL='GRADE LEVEL RECEIVING INSTRUCTION MATHEM'
X014501  SCIGRADL LABEL='GRADE LEVEL RECEIVING INSTRUCTION SCIENC'
X014801  RDGPARTL LABEL='HOW PARTICPATE IN NAEP READING LANGUAGE '
X015001  SCIPARTL LABEL='HOW PARTICIPATE IN NAEP SCIENCE/LEP     '
```

Table A20. Value Labels for Variables Used in Linear Regressions

```
VALUE ACCOMTY   1='BIL BK/SCI GLOSSARY '    2='BILINGAL DICTIONARY '
                3='LARGE PRINT BOOK    '    4='EXTND TIME REG SES  '
                5='READ ALOUD          '    6='SMALL GROUP         '
                7='ONE-ON-ONE          '    8='SCRIBE/USE OF PC    '
                9='OTHER/SPECIFY       '   ;
VALUE X012101Q  0='MULTIPLE RESPONSE   '    1='SPECIFIC LEARNING   '
                2='HEARING IMPAIRMENT  '    3='VISUAL IMPAIRMENT   '
                4='SPEECH IMPAIRMENT   '    5='MENTAL RETARDATION  '
                6='EMOTIONAL DISTURBANC'    7='ORTHOPEDIC IMPAIRMNT'
                8='BRAIN INJURY        '    9='AUTISM              '
               10='DEVELOPMENTAL DELAY '   11='OTHER HEALTH        '
               12='OTHER               '   88='OMITTED             ';
VALUE X012201Q  0='MULTIPLE RESPONSE   '    1='PROFOUND/SEVERE     '
                2='MODERATE            '    3='MILD                '
                8='OMITTED             '   ;
VALUE X012401Q  0='MULTIPLE RESPONSE   '    1='NOT RECEIVING INSTRU'
                2='AT OR ABOVE GR LEVEL'    3='1 YR BELOW GR LEVEL '
                4='2+ YRS BELOW GR LEV '    7='I DON''T KNOW        '
                8='OMITTED             '   ;
VALUE X012501Q  0='MULTIPLE RESPONSE   '    1='NOT RECEIVING       '
                2='YES                 '    3='NO                  '
                7='I DON''T KNOW        '   8='OMITTED             ';
VALUE X012601Q  0='MULTIPLE RESPONSE   '    1='NOT RECEIVING       '
                2='AT OR ABOVE GR LEVEL'    3='1 YR BELOW GR LEVEL '
                4='2+ BELOW GR LEVEL   '    7='I DON''T KNOW        '
                8='OMITTED             '   ;
VALUE X013001Q  0='MULTIPLE RESPONSE   '    1='YES                 '
                2='NO                  '    3='STUDENT CANNOT TEST '
                8='OMITTED             '   ;
VALUE X013501Q  0='MULTIPLE RESPONSE   '    1='WITHOUT ADAPTATIONS '
                2='WITH ADAPTATIONS    '    3='STUDENT CANNOT TEST '
                8='OMITTED             '   ;
VALUE X013801Q  0='MULTIPLE RESPONSE   '    1='SPANISH             '
                2='ANOTHER LANGUAGE    '    8='OMITTED             ';
VALUE X013901Q  0='MULTIPLE RESPONSE   '    1='NOT IN ENGLISH      '
                2='1 YEAR              '    3='2 YEARS             '
                4='3 YEARS             '    5='4 YEARS OR MORE     '
                7='I DON''T KNOW        '   8='OMITTED             ';
VALUE X014201Q  0='MULTIPLE RESPONSE   '    1='0%                  '
                2='1-24%               '    3='25-49%              '
                4='50-99%              '    5='100%                '
                8='OMITTED             '   ;
VALUE X014301Q  0='MULTIPLE RESPONSE   '    1='NOT RECEIVING INSTRU'
                2='AT OR ABOVE GR LEVEL'    3='1 YR BELOW GR LEVEL '
                4='2+ BELOW GR LEVEL   '    7='I DON''T KNOW        '
                8='OMITTED             '   ;
VALUE X014801Q  0='MULTIPLE RESPONSE   '    1='ENGLISH W/NO ADAPTS '
                2='ENGLISH W/ADAPTS    '    3='NATIVE LANGUAGE     '
                4='NOT  PARTICIPATE    '    8='OMITTED             ';
```

Table A21. Recodings of Variables Used in Linear Regressions

```
lrndisab=(x012101 in (1,5,8,9));

if x012201 not in (1,2,3) then x012201=4;
if x012401 not in (3,4)   then x012401=2;
if x012501 not in (3)     then x012501=2;
if x012601 not in (3,4)   then x012601=2;
if x012701 not in (3)     then x012701=2;
if x012801 not in (3,4)   then x012801=2;
if x012901 not in (3)     then x012901=2;
if x013501 not in (2,3)   then x013501=1;
if x013701 not in (2,3)   then x013701=1;
if x013901 not in (1,2,3,4) then x013901=5;
if x014001 not in (1,2,3,4) then x014001=5;
if x014101 not in (1,2,3,4) then x014101=5;
if x014201 not in (2,3,4,5) then x014201=1;
if x014301 not in (3,4)   then x014301=2;
if x014401 not in (3,4)   then x014401=2;
if x014501 not in (3,4)   then x014501=2;
if x014801 not in (2,3,4) then x014801=1;
if x015001 not in (2,3,4) then x015001=1;
```

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

NOTICE

Reproduction Basis

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

EFF-089 (1/2003)